

## Historical social research: the use of historical and process-produced data

Clubb, Jerome M. (Ed.); Scheuch, Erwin K. (Ed.)

Veröffentlichungsversion / Published Version  
Sammelwerk / collection

Zur Verfügung gestellt in Kooperation mit / provided in cooperation with:  
GESIS - Leibniz-Institut für Sozialwissenschaften

### Empfohlene Zitierung / Suggested Citation:

Clubb, J. M., & Scheuch, E. K. (Eds.). (1980). *Historical social research: the use of historical and process-produced data* (Historisch-Sozialwissenschaftliche Forschungen : quantitative sozialwissenschaftliche Analysen von historischen und prozeß-produzierten Daten, 6). Stuttgart: Klett-Cotta. <https://nbn-resolving.org/urn:nbn:de:0168-ssoar-326045>

### Nutzungsbedingungen:

Dieser Text wird unter einer Deposit-Lizenz (Keine Weiterverbreitung - keine Bearbeitung) zur Verfügung gestellt. Gewährt wird ein nicht exklusives, nicht übertragbares, persönliches und beschränktes Recht auf Nutzung dieses Dokuments. Dieses Dokument ist ausschließlich für den persönlichen, nicht-kommerziellen Gebrauch bestimmt. Auf sämtlichen Kopien dieses Dokuments müssen alle Urheberrechtshinweise und sonstigen Hinweise auf gesetzlichen Schutz beibehalten werden. Sie dürfen dieses Dokument nicht in irgendeiner Weise abändern, noch dürfen Sie dieses Dokument für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, aufführen, vertreiben oder anderweitig nutzen. Mit der Verwendung dieses Dokuments erkennen Sie die Nutzungsbedingungen an.

### Terms of use:

This document is made available under Deposit Licence (No Redistribution - no modifications). We grant a non-exclusive, non-transferable, individual and limited right to using this document. This document is solely intended for your personal, non-commercial use. All of the copies of this documents must retain all copyright information and other information regarding legal protection. You are not allowed to alter this document in any way, to copy it for public or commercial purposes, to exhibit the document in public, to perform, distribute or otherwise use the document in public. By using this particular document, you accept the above-stated conditions of use.

102 Cluj 1

**HSF**



# **Historisch-Sozialwissenschaftliche Forschungen**

Quantitative sozialwissenschaftliche Analysen  
von historischen und prozeß-produzierten Daten

Herausgegeben von  
Heinrich Best, Wolfgang Bick  
Reinhard Mann, Paul J. Müller  
Herbert Reinke, Wilhelm H. Schröder

Zentrum für historische Sozialforschung

Band 6

Klett-Cotta

Jerome M. Clubb, Erwin K. Scheuch (eds.)

## **Historical Social Research**

The Use of Historical  
and Process-Produced Data

Klett-Cotta

**CIP-Kurztitelaufnahme der Deutschen Bibliothek**

**Historical Social Research: The Use of Histor. and Process-Produced Data/  
Jerome M. Clubb; Erwin K. Scheuch (eds.). — Stuttgart: Klett-Cotta, 1980.  
(Historisch-sozialwissenschaftliche Forschungen; Bd. 6)**

**ISBN 3-12-911060-7**

**NE: Clubb, Jerome M. [Hrsg.]**

Alle Rechte vorbehalten

Fotomechanische Wiedergabe nur mit Genehmigung des Verlages  
Verlagsgemeinschaft Ernst Klett — J. G. Cotta'sche Buchhandlung  
Nachfolger GmbH

© Ernst Klett, Stuttgart 1980. Printed in Germany

Satz: Composersatz U. Sommer, Eutingen

Druck: Röck, Weinsberg

# Contents

Preface		9
I. The Emergence of Social Scientific History and the Analysis of Process-Produced Data		
Jerome M. Clubb	The „New“ Quantitative History: Social Science or Old Wine in New Bottles?	13
Erwin K. Scheuch	Quantitative Analysis of Historical Material as the Basis for a New Cooperation Between History and Sociology	25
Wolfram Fischer	The Emergence of Social Scientific History and the Analysis of Process-Produced Data: Some Introductory Remarks	47
II. Analysis of Administrative and Census Data		
Maris A. Vinovskis	Problems and Opportunities in the Use of Individual and Aggregate Level Census Data	53
Arthur E. Imhof	Reconstructing Biological Frameworks of Populations in the Past	71
Theodore Hershberg	Interdisciplinary Research at the Philadelphia Social History Project: Analytic Goals, Data and Data Manipulation Strategies for the Study of the Nineteenth-Century Industrial City	84
Erdmann Weyrauch	Conditions and Chances in the Computer-Aided Analysis of Historical Record Series: The Case of Early Modern Urban Administration	112
Horst Matzerath	Aggregated Census Data and the Analysis of Urbanization in Prussia (1816–1939)	120
III. Analysis of Life Histories		
Karl-Heinz Reuband	Life Histories: Problems and Prospects of Longitudinal Designs	135
Walter Müller	The Analysis of Life Histories: Illustrations of the Use of Life History Plots	164

Göran Sandström Jan Sundin	Computer Analysis of Life Histories from Swedish Church Records: A Case Study from the Demographic Data Base at Umea University	192
Wilhelm Heinz Schröder	Quantitative Analyses of Collective Life Histories: The Case of the Social Democratic Candidates for the German Reichstag 1898–1912	203
Walter M. Pinter	The Use of Collective Biography in Research on the Imperial Russian Civil Service	225

#### IV. Content and Document Analysis

Ekkehard Mochmann	Computer Aided Content Analysis of Historical and Process-Produced Data: Methodological and Technical Aspects	235
Heinrich Best	Analysis of Content and Context of Historical Documents — The Case of Petitions to the Frankfurt National Assembly 1848/49	244

#### V. Time Series Analysis

Jerald Hage Edward T. Gargan Robert Hanneman	Procedures for Periodizing History: Determining Eras in the Histories of Britain, France, Germany and Italy	267
Aage B. Sørensen	Analysis of Change in Discrete Variables	284
Thomas Kuczynski	Have There Been Differences Between the Growth Rates in Different Periods of the Development of the Capitalist World Economy Since 1850? An Application of Cluster Analysis in Time Series Analysis	300

#### VI. Network Analysis

Frans N. Stokman Wijbrandt H. van Schuur	Leadership Dimensions Among Developing Nations in the United Nations	319
Frans N. Stokman	Graph Theoretical Elaboration of Cumulative Scaling Techniques	342
Robert J. Mokken	Cliques, Clubs and Clans	353

## VII. Structure of Social Bookkeeping: Problems with the Generation of Complex Data Files

Wolfgang Bick	The Nature of Process-Produced Data —	369
Paul J. Müller	Towards a Social-Scientific Source Criticism	
Ian Winchester	Priorities for Record Linkage: A Theoretical and Practical Checklist	414
John de Vries	Problems in Handling Process-Produced Data	431
Philip J. Stone	A Perspective on Social Science Data Management	444

## VIII. Preservation, Storage, and Access

Charles M. Dollar	Problems and Procedures for Preservation and Dissemination of Computer-Readable Data	457
Wolf Buchmann	Archives and Machine-Readable Data from Public Administration in the Federal Republic of Germany	473
Per Nielsen	How To Teach Data Procedures „The Noble Art“ of Data Documentation	477

## IX. New Data Bases

Charles Lewis Taylor	Development of Cross National Comparative Data from Social Bookkeeping Sources	491
Herbert Reinke	The Analysis of Change and Persistence in German Society: The German Census of Occupations as a New Data Base	501
Jarle Brosveet	Standardization of Longitudinal, Aggregate-Level Data in the Norwegian Commune Database	513
Peter Flora	HIWED — A Comparative Historical Research Project on Western Europe	524

Contributors		533
--------------	--	-----



## Preface

This volume explores prospects and problems of a type of historical social research that is not yet routine but already important: Quantitative historical social research. Sociology and history were intimately linked much earlier in the development of both disciplines. This is then a renewed encounter between the two disciplines and in many ways a more complete interpenetration than before. The volume contains contributions on methodological aspects, theoretical implications, and infrastructural preconditions of the renewed encounter between history and sociology.

An intensified exchange and even a rapprochement between both sciences can be observed during the past decade or more. This development has its origin in a re-orientation of both history and sociology: The traditional interest of historical research in individual motivations and actions, personalities and events is complemented by an interest in the structural preconditions and effects of collective behavior and attitudes in past times. This „collective history“ requires quantification as an adequate methodological base.

In a parallel development many sociologists share a rising discomfort with the restricted scope, limited validity and diminishing reliability of survey data. A more extensive use of historical and process-produced data is one reaction to this reappraisal. There is also an obvious tendency to resume in a modified way the evolutionary and historical traditions of sociological theory-building.

In view of these developments we can observe more and more correspondences and parallels between sociology and history in terms of data bases, methods, and theoretical conceptions. This seemed an opportune moment to complement and intensify written communications through direct communications between scientists of both branches that only a conference could afford. Thus QUANTUM initiated and organized the international and interdisciplinary conference on „Quantification and Methods in Social Science Research: possibilities and problems with the use of historical and process-produced data“. It was held at the University of Cologne, on August 10–12, 1977. For the first time in Europe, this event brought together historians, sociologists, political scientists, geographers, computer scientists, and personnel from archives and data organizations to demonstrate and to discuss their (possible) contributions to a quantitative historical social research. This volume includes a selection of the papers presented then. All of them were originally written for the conference. The papers were somewhat revised for publication.

The book covers a wide range of topics and documents a variety of methodological approaches typical of historical social research. Some of the contributions, presenting projects in very initial stages, may not meet the stricter criteria that are appropriate for more „mature“ fields of research. However, after some consideration the editors preferred to document the heterogeneity of issues and the different stages of development in historical social research instead of applying rigorous criteria for selection. In this way the volume represents the current state of the art as it happens to be, and may in this way influence further research in the field.



The conference — and in consequence this volume — would not have been possible without the financial support of the Deutsche Forschungsgemeinschaft (German Science Foundation) and the extensive work that was done by the QUANTUM executive board — Heinrich Best, Wolfgang Bick, Reinhard Mann, Paul J. Müller, Herbert Reinke, Karl-Heinz Reuband, Wilhelm H. Schröder — in conceptualizing and organizing the meeting. The participation of American scholars at the conference was made possible by the Social Science History Association and financially supported by the National Science Foundation, USA.

Finally we wish to express our gratitude to Heinrich Best, Center for Historical Social Research, for promoting this volume.

Cologne  
February, 1979

Jerome M. Clubb  
Erwin K. Scheuch

# I. The Emergence of Social Scientific History and the Analysis of Process-Produced Data



The „New“ Quantitative History: Social Science  
or Old Wine in New Bottles?

Much has been written and said in recent years of the emergence of a variety of „new“ histories<sup>1</sup>. To the casual observer the diversity of the new histories may seem bewildering and even convey the impression of pretentious craving for the appearance, but not the substance, of innovation. And, indeed, the new histories have provoked substantial controversy and criticism on exactly those grounds. Yet even the most cursory examination of the recent literature suggests numerous departures from the preoccupations of traditional historiography. New themes are explored, new methods employed, new sources exploited and neglected social institutions examined in new ways. Some of the elitist bias that has long been a dominant characteristic of historical research and writing has dissipated as historians have attempted to use new sources to examine the conditions and behavior of the ordinary people of the past. Above all, perhaps, a growing number, but still a minority, of historians have made increased use of quantitative methods and tools. If there is a common denominator that cuts across the varieties of the new histories, it is quantification, and it is possible to speak with accuracy of the recent emergence of a quantitative form of historical inquiry.

These new scholarly interests and endeavors have been marked by their fair share of false starts and mistaken directions. A cadre of curmudgeonly critics among historians, and even including a few practitioners of the other social sciences, have declared all such efforts barren, not history, and fated only to do violence of the literary values and the humane traditions of conventional historiography. It is certainly true that claims for the new histories have sometimes been unduly exalted and that the new approaches to the study of the past have not yet produced the revolutionary new knowledge that their more optimistic and aggressive proponents sometimes promised. Yet it is also true that these new efforts have already enriched historical studies.

The purpose here, however, is not to detail or celebrate the characteristics, accomplishments and failures of each of the new histories. To do so would require a treatise of tiresome length. In any event, numerous „state of the art“ essays con-

<sup>1</sup> This paper draws upon two essays written in collaboration with Professor Allan G. Bogue and William H. Flanigan and published in *American Behavioral Scientist*, 21 (November/December 1977). While the present paper profits greatly from their erudition and insights, neither of them is responsible for errors of fact, judgment or interpretation which it may include.

cerned in one way or the other with these new approaches to historical inquiry have appeared in recent years<sup>2</sup>. These essays, and the numerous works which they reference, amply demonstrate both the diversity and the vitality of the new histories. The goals of the present essay are different and more modest. It is concerned with quantitative approaches to historical inquiry primarily as practiced by historians in the United States and, in the first instance, with progress toward the emergence of a social scientific history — or, more properly, with progress toward a form of historical inquiry devoted to the development of scientific knowledge of human behavior. A second concern is with obstacles that lie in the way of a form of historical inquiry so conceived and motivated.

### The „New“ Quantitative History

Tables, graphs, charts and even elaborate statistical formulations have become common elements in the research reports of historians. Quantification has become an accepted, although by no means universally welcomed, element of the historical discipline. The use of quantitative methods and materials by historians has opened the way for forward steps in the advancement of historical knowledge; major contributions to historical knowledge that rest in critical ways on the application of quantitative tools and materials can be cited; and the use of those tools and materials has resulted in a literature that often appears quite unlike the product of more conventional historical research.

In view of these obvious manifestations of change within the historical discipline it may seem remarkable that, from many perspectives, the most striking developments in historical inquiry have come not from professional historians but from scholars outside the discipline. And here, perhaps to betray a parochial bias, recent comparative studies of political and societal change — whether termed modernization, development or described under some other rubric — that sweep broadly across both modern history and the contemporary era might be particularly singled out<sup>3</sup>. The ultimate value of the theoretical and conceptual formulations that have emerged from such studies is surely questionable, but it is certainly the case that they cast the politics and society of the past in new perspective, point new directions for historical research, and call into doubt long-accepted explanations of his-

<sup>2</sup> See, for example, Bogue, Allan G. (ed.), *Emerging Theoretical Models in Social and Political History*, in: *American Behavioral Scientist*, 16 (May/June 1973); Bogue, Allan G., and Clubb, Jerome (eds.), *History and the Social Sciences: Progress and Prospects*, in: *American Behavioral Scientist*, 21 (November/December 1977); and Lorwin, Val R., and Price, Jacob M. (eds.), *The Dimensions of the Past: Materials, Problems and Opportunities for Quantitative Work in History*, New Haven 1972. See also Swierenga, Robert P., *Computers and American History: The Impact of the „New“ Generation*, in: *The Journal of American History*, 60 (March 1974); and *The New Political History: Progress and Prospects*, in: *Computers and the Humanities*, 11 (September/October 1977).

<sup>3</sup> This is not to overlook the pioneering work of such historians as Cyril E. Black.

torical events. They attempt as well to present an integrated view of politics and society that is often — perhaps usually — missing from the work of historians.

The historical studies conducted by social scientists are remarkable enough in their own right. Two considerations seem more remarkable. One is the fact that social scientists have not placed greater reliance upon the work of historians but have tended instead to develop their own form of historical inquiry. A second and, for present purposes, more important consideration is the reactions of historians to these studies, at least in the United States, and including many quantitative historians. For many historians, both „new“ and otherwise, the scope of such largescale comparative studies seems monumental, even grandiose. Critical relations seem to be merely assumed rather than demonstrated; conceptual and theoretical formulations seem excessively abstract and mechanistic and not easily operationalized in terms of historical data. Above all, perhaps, such studies seem to play too fast and loose with chronology and to lack the rich sense of time, place and specific context to which historians are accustomed. Seemingly disparate phenomena are classified, lumped together and compared apparently with little regard for temporal order or historical context. These considerations suggest differences in orientation between historians, on the one hand, and social scientists who use the past as a source of evidence, on the other. They suggest, as well, that history, even quantitative history, has moved only a little way in the direction of social science.

The basis of these differences and their significance can be partially elucidated by brief consideration of the genesis of quantitative approaches to historical inquiry. It is frequently suggested that quantitative history is characterized by a marked affinity with the related social sciences and that, indeed, these new approaches grew out of the social sciences. In fact, affinity with the social sciences is clear and explicit in terms of the use of quantitative techniques and source materials. An affinity is also present, but substantially less explicit and less clear, where conceptual and theoretical formulations are concerned. But the new quantitative history can also and perhaps more accurately be seen as a product of basic canons of traditional historical method. To a considerable degree, the new use of quantitative techniques and source materials grew out of the traditional historiographical requirement that all source materials and all relevant methods be used to gain the most detailed, complete, and objective view possible of past events, personages and situations. Put differently, the use of new techniques and source materials has meant no necessary deviation from the goal of traditional historiography — to describe and reconstruct the past „as it actually happened“. Neither does that use involve necessarily a departure from the methodological assumptions of traditional historiography, nor does it necessarily look directly toward development of scientific knowledge of human affairs.

These considerations may help to explain some of the characteristics of early applications of quantitative techniques and materials by historians. In general, quantitative historians did not escape the tendency to focus research upon limited historical episodes and problems. That tendency in turn also worked, of course, to promote an excessively compartmentalized and static view of human affairs. If any-

thing, moreover, the higher costs of research using quantitative data and methods, as compared with more conventional approaches, led initially to concern for even more narrow and more limited research topics and to heavier emphasis upon case studies. Thus the findings produced by much of the early work were of limited or, at best, unknown generality. This is not to issue a blanket indictment of the case study approach. It is to say, however, that the early view — that multiplication of case studies in the absence of comparable methods and data and without theory would lead to general findings — was fallacious.

A more serious legacy of traditional historiography was in the areas of concept and theory. As David Potter observed in 1963:

Orthodox or formal historical method was shaped at a time when men believed that a body of data would reveal its own meaning and would interpret itself, if only it were valid or authentic and were arranged in time sequence. The central problem of method, therefore, was to validate the data rather than to interpret them. Hence the problem of historical interpretation was neglected; indeed its very existence as a problem was denied at the theoretical level, and the principal questions which the problem of interpretation ought to have posed were left to non-historians. Thus the problem of causation has been left to the philosophers; the problem of human motivation has been left to psychologists; the problem of social organization has been left to the sociologists<sup>4</sup>.

Thus valid understanding of historical events, processes and persons could be gained simply by consulting all relevant sources — or, through „emersion in the sources“ as it is still often put. Causal relations were seen as self-evident or asserted on the basis of intuition; human attitudes and motivation were also seen as self-evident or identifiable essentially through processes of empathy. Interpretation, systematic theory and generalization were unnecessary. To a considerable degree the early quantitative historians retained these views and assumptions, and it is by no means the case that they have yet entirely disappeared.

The impact of these assumptions was readily observable in much of the early work of quantitative historians. Energy was invested in acquiring conversance with the tools of quantification. Emphasis was placed, however, on techniques of data analysis, and for many data analysis was apparently simply an alternative means to „emersion in the sources“. But despite investment of energy, familiarity with techniques of data analysis was usually gained through self-training and in essentially „cookbook“ fashion. Outside the field of economic history, neither training opportunities, the curriculum in history, nor the structure of professional rewards were such as to encourage acquisition of meaningful formal training in mathematics and related areas. The consequence was that the properties of statistical procedures were often not fully understood with the further result that all too frequently applications of those procedures were naive and erroneous.

Too little attention was directed to problems of conceptualization, measurement

<sup>4</sup> Potter, David M., *Explicit Data and Implicit Assumptions in Historical Study*, in: Gottschalk, Louis (ed.), *Generalizations in the Writing of History: A Report of the Committee on Historical Analysis of the Social Science Research Council*, Chicago 1963.

and inference. It was not fully recognized that selection of appropriate statistical models requires conceptualization of the historical processes and phenomena of concern. In the absence of that conceptualization, inappropriate statistical models were frequently applied. Much of the early work of quantitative historians involved a marked tendency to, in effect, „reify“ empirical data and to treat empirical data as effectively synonymous with underlying concepts. The measurement gap — the gap between data and concept — was overlooked. Thus voting records were sometimes treated as direct measures of mass partisan attitudes or popular policy preferences, occupational or educational characteristics were taken as straightforward indications of social status or position, and change in such characteristics was taken as a direct measure of social mobility. The need for conceptualization — for auxiliary theory — to link data to concept was not recognized. Causal inferences were often drawn, asserted and refuted, but without adequate empirical warrant or test and without theoretical justification<sup>5</sup>.

The early forays into quantification, in short, did not free historians from the assumption of nineteenth century „scientific“ historiography that data are somehow self-interpreting. Like their more conventional colleagues, quantitative historians tended to be essentially atheoretical, although important exceptions could be cited. It was not, of course, that theoretical formulations and generalizations were absent from their work. In fact, the actual structure of explanation in numerous quantitative historical works was provided, not by historical sources, but by generalizations and theories — by „covering laws“ — drawn from common sense, personal philosophy, ideology, or an obsolete social science. Indeed, without these generalizations and theoretical formulations, meaningful and explanations of historical events could not have been constructed<sup>6</sup>. These formulations, however, were usually implicit and untested and were seldom subjected to self-conscious and critical examination.

The new economic, or „cliometric“, history — theoretically and mathematically the most elegant and advanced of the new histories — was a partial exception. In this area neo-classical economic theory was explicitly and rigorously employed as a basis for estimation and measurement, to interpret and „make sense“ of historical data, and to resolve long-contested questions in economic history. On the other hand, neoclassical theory was primarily used to explain the phenomena of the past; substantially less attention was directed to using the data of the past to extend, test and refine economic theory<sup>7</sup>.

<sup>5</sup> For a discussion of one illustration of these problems see Clubb, Jerome M., and Allen, Howard W., *Collective Biography and the Progressive Movement: The 'Status Revolution' Revisited*, in: *Social Science History*, 1 (Summer 1977).

<sup>6</sup> The point of view expressed here draws upon Hempel, C. G., *The Function of General Laws in History*, in: Gardiner, Patrick (ed.), *Theories of History*, Glencoe/Ill. 1959; Murphey, Murray G., *Our Knowledge of the Historical Past*, Indianapolis/Ind. 1973; and Potter, *Explicit Data and Implicit Assumptions*.

<sup>7</sup> See North, Douglass C., *The New Economic History after Twenty Years*, in: *American Behavioral Scientist*, 21 (November/December 1977).



Substantial progress has been made in many of these respects. Greater attention is now given to matters of representation. The case study approach is less commonly employed, and historians less frequently claim generality for findings based upon case studies. Proficiency in the use of analytical techniques has grown, and an increasing number of historians have benefitted from formal training in the tools and methods of quantitative inquiry. In the use of these tools historians have become more sophisticated and subtle, and at least passing concern is directed to problems of measurement. Thus it is possible to speak of the emergence of a new history. Historians now attempt to gain a more complete view of the past; they are more comprehensive in their use of source material; and they are no longer satisfied with a history that merely recounts the doings of the few of power and position of the past. Rather historians have devoted greater attention to the ordinary men and women of the past, to basic but neglected social institutions, and to events and conditions at the „grass roots“.

But if we speak of central tendencies, this is a history that remains primarily dedicated to traditional historiographical goals. While historians have become more attuned to the related social sciences, their use of generalization and theory is still largely implicit and uncritical, and the role of theory and generalization in historical explanation is as yet only partially recognized. The quantitative work of historians is primarily concerned with specific episodes and events which are selected for examination on the basis of their intrinsic historical interest. In the main, the goal of quantitative historians is to reconstruct and describe the past „as it actually happened“, and reconstruction and description of past events is taken as an end in itself. On the other hand, substantially less in the way of systematic and self-conscious effort is directed to using the evidence of the past to develop generalizations and to refine and construct theoretical formulations. Whether pursued in quantitative or other ways, these goals and efforts fall short of social science and do not promise to fully realize the potentiality of historical inquiry as a means to contribute to scientific knowledge of human affairs.

It may well be that recognition of the value of quantitative methods and materials as tools of historical inquiry will lead to reorientation of the historical discipline in social scientific directions. Quantification is not, however, the only attribute of science. As yet, the use of quantitative methods and materials has led few historians, to use Charles Tilly's words, „quite outside history, in[to] that timeless realm in which situations, persons, or events plucked from the past or the present serve as tests of general statements about social life.“<sup>8</sup> A new quantitative history has emerged, but that history carries us only a little way toward contribution to the development of scientific knowledge of human affairs.

This is not to say, of course, that all historians will, or should, pursue social scientific goals. The effort to develop valid and reliable descriptions and explana-

<sup>8</sup> Tilly, Charles, *Quantification in History, As seen from France*, in: Lorwin, *The Dimensions of the Past*, p. 108.

tions of historical events is obviously a laudable intellectual enterprise in its own right. Indeed, valid and reliable reconstruction of past events and situations would be a vital component of social scientific historical inquiry. It is to say, however, that historical inquiry dedicated to scientific ends would require a further and more radical transformation than quantitative, or other, historians have yet accomplished. It is also to say that because of the central and necessary role of generalization and theory in historical explanation, elements of that transformation are required if more valid and reliable reconstruction of past events is to be achieved.

## Problems of Historical Data

Aspects of the required transformation, and broad characteristics of the form of historical studies that would be its consequence, can be readily recognized if the goals and limitations of the contemporary social sciences are considered. A central goal of the social sciences is to identify regularities in human affairs and to develop empirically refutable theoretical formulations which link together and explain those regularities. In the pursuit of this goal, however, the social sciences are constrained by sharply limited capacity to experiment where human beings are concerned. To compensate for these limitations, social scientists have recognized the necessity of examining and comparing social phenomena in a variety of situational contexts, and attention is devoted to comparative studies across nations, cultures, regions and subnational groups. The past affords an opportunity to extend these efforts and to examine a wider variety of human behavior in a wider variety of contexts.

Viewed in these terms, the past can be used to formulate hypotheses and to test those based upon investigation of contemporary phenomena, to examine developmental processes, to systematically trace trends and change over time, and to identify the determinants of societal change. The task of a genuinely social scientific historian would be to use the past to construct empirical social theory rather than to merely use social theory to describe and explain specific events of the past. To do so effectively, however, would require that historians give greater attention to the epistemological bases and limitations of their work. Greater and more serious attention to matters of method would also be required and not merely to matters of technique but to broader issues of measurement, design, conceptualization and inference as well. In their teaching and research historians would devote less attention to the specific facts and episodes of history and more to interpretation and theory. Substantial attention might also be directed to classification of historical and contemporary events, processes, institutions and populations in terms of properties relevant to particular theoretical formulations. By the same token, specific events and phenomena would be investigated not for their intrinsic historical interest but in terms of their relevance to theoretical concerns.

But even if reorientations of this sort are both possible and desirable, critical problems are encountered. To a considerable degree these problems relate to the characteristics and limitations of historical source materials, and, it is worth nothing,

these problems are equally severe where achievement of the goals of traditional historiography are concerned. Indeed, in view of the gravity of these problems it is legitimate to ask whether historical inquiry can ever be any more than a tincture of empirical evidence combined with bits of useful theory and mixed with large elements of impression, surmise and empathetic understanding.

The source materials upon which historians must rely are virtually by definition „process produced.“ It is true that historians occasionally have at their disposal data that were collected by social scientists to serve the purpose of social scientific research. As the specious present slides imperceptibly into the historical past, we can imagine that historians will turn to the rich sample survey data now being religiously collected and maintained by social science data archives. But we can also imagine that historians in the future will regard these data as no less process produced — produced in this case by the process of social research as archaically practiced in the mid-twentieth century — and will bemoan the fact that the wrong data were collected, the wrong questions asked, and that underlying assumptions and methods were not better documented.

The central difficulty where process-produced data are concerned is, of course, that the data were not collected, compiled and preserved with the needs of researchers in mind. Rather they are simply the byproducts of ongoing social, governmental and economic processes. Historical process-produced data involve additional complexities for these are data that have been filtered and winnowed by the processes of time. Historical research, in other words, is limited not only to data — and information about data — that was once collected but also to data and information that have survived. Historical data, in other words, are the residual process-produced data that have survived the ravages of time.

In their work historians sometimes confront an embarrassment of riches but most frequently their research must be conducted below the data poverty line. For a few problems, relevant data are voluminous indeed, and their effective use would involve a truly monumental data collection and processing effort. For most problems, however, data are at best incomplete, and for vast areas of historical concern data are virtually nonexistent. Aside from a few treasure troves, for example, data that bear on mass attitudes in the past do not exist, and statements in this area must be based on often dubious inferences from behavioral data.

Historians, of course, give great thought and attention to their sources. Indeed, source criticism is a central element of traditional historical methods. But conventional source criticism carries us only a little way. External source criticism is concerned with assessment of the authenticity of sources; internal criticism involves examination of texts to identify error, misrepresentation and inconsistency. A systematic historical inquiry that attempts to conform to scientific canons of verification and explanation requires that much more be known of the nature and properties of sources and data.

Murray Murphy is one of the few historians who has devoted systematic attention to the characteristics of historical data and to the methods available for their use. He singles out five methodological problems that result from the characteristics

of historical data<sup>9</sup>. While these problems could be summarized in various ways, Murphey's discussion is both succinct and useful. The first of these is the problem of quantity which was just alluded to. The second is the problem of aggregation. Historical data are frequently available at a level of aggregation that is inappropriate to research concerns. In the most usual case, data are available in the form of summary measures or indices or as aggregated values for population groups and geographical units. It is not, of course, that summary measures and aggregated data are useless for many of the purposes of research. In fact, such data are frequently useful indeed. It is only that a variety of research problems require disaggregated data. A third problem is that of sampling. Most — in a sense all — historical data are incomplete and can be seen as samples from total universes. In most cases, however, neither the properties of the universe nor of the sample are known. Murphey describes a fourth problem as that of „informant bias“ which, put simply, is the problem of the reliability and, unlike contemporary social scientists, the historian has no opportunity to query and cross examine informants to determine reliability.

The final problem identified by Murphy is that of measurement which is part and parcel of the general problem of measurement confronted by all social scientists. In the historian's case, however, the problem is enormously complicated by the fragmentary nature of historical data and by the absence of data that bear upon the perceptions, the attitudes, and the cognitive systems of historical individuals. To these problems a sixth, that of specification, can be added, although it is really subsumed under other categories. The simple fact is that historians frequently cannot adequately specify models of historical processes or effectively test for spurious relations because appropriate data were either never collected or were not preserved. As a consequence, capacity to convincingly refute or support otherwise meaningful hypotheses is often sharply limited.

Again, these problems could be stated in a variety of ways, and the preceding summary may do some violence to Murphey's formulations. Even so it may serve to point the issues. We can note that Murphey's formulations seem to reflect an undue preoccupation with research that focuses upon individuals. Research at a more macroscopic level may not confront these same problems, or, at least, not with the same severity. And, as Murphey points out, some of these problems may be amenable to at least partial solution. Even at best, however, the critique is chastening. Clearly, a systematic history that pays due regard to problems of data, verification and inference cannot provide as rich or as fullsome a view of the past as is often pretended to through intuitive, empathetic and impressionistic approaches. It is clear as well that capacity to use the past as a source of evidence to test and construct social theory is similarly limited. Perhaps the first lesson to be learned is one of limitations.

This is not to suggest, however, that the limitations and fragmentary nature of historical data preclude, as has sometimes been argued, application of advanced and complex quantitative techniques. Indeed, quite the reverse is the case. The very in-

<sup>9</sup> Murphey, *Our Knowledge of the Historical Past*, especially Chapter 6.

adequacies of historical data require more complex analytical techniques and more subtle and sophisticated applications than are required for more perfect and less obstreperous data. If anything, moreover, greater methodological expertise is probably required for fully effective research use of historical data than is required for much of the more perfect contemporary data employed by social scientists.

Certainly the limitations of historical data dictate research in new areas. Methodological research of the sort that Murphey suggests is obviously required. Areas in which methodological innovation is clearly needed are also obvious. They include techniques for sampling from incomplete records, for estimating missing data, and, most obviously, for estimating individual level relations from aggregated data. And here the picture is even more chastening. While the technical expertise of historians has steadily risen in recent years, it is still, at least in the United States, low indeed compared with that of the related social sciences. Few historians are equipped to carry out methodological research or to solve methodological problems, and unfortunately, given the nature of training programs in the universities of the United States, opportunities and inducements to acquire necessary expertise are still by no means abundant. Thus, there is need for the development of alternative sources of training including that of a „retooling“ nature.

And more can be done to improve capacity to employ historical data and to facilitate better realization of the potentialities of the evidence of the past for the pursuit of social scientific goals. Accomplishment of these tasks would require that historians place greater value on prosaic but demanding activities that are now neither particularly well-regarded nor well-recognized in the distribution of professional rewards. Substantial efforts might usefully be directed to the development of systematic measures and indicators of social phenomena extending across long historical periods. These tasks would involve more than data collection; rather they would require systematic evaluation of data sources and collection and combination of data to create estimates to measure and reflect substantive and theoretical concepts. Examples of work of this sort include the historical economic indexes and estimates developed by the National Bureau of Economic Research, the efforts of Walter Dean Burnham, and the Inter-University Consortium for Political and Social Research to create reliable estimates of mass electoral participation and behavior for the United States during the nineteenth and twentieth centuries, as well as the systematic information on the characteristics of the international state system during the same period developed by J. David Singer and his associates.

Indeed, a new form of historiography is needed, as Charles Tilly suggests in another context, one that is concerned with the nature and properties of historical sources<sup>10</sup>. Much more can be learned of the structure of social bookkeeping systems of the past. If more was known of the functions of historical administrative systems, for example, improved capacity to estimate the kinds of biases and error

<sup>10</sup> Tilly, *Quantification in History*, pp. 110 *passim*. See also Narroll, Raoull, *Data Quality Control*, New York 1970.

that are likely to characterize the data which they gathered would be gained. Most historians have encountered shifts in time series which might signify change in production levels, in the incidence of violence, or in rates of voter participation but which might also reflect no more than change in recording procedures or in definitions or classification systems. Better knowledge of the operation of social bookkeeping systems might provide clues for the interpretation of such changes. And much the same can be said of information preservation. If more was known of the functions that information preservation was intended to serve in past situations, then historians might be better equipped to estimate the representative quality and the limitations of surviving records.

Inquiry of this sort would not be limited to the formal operations and functions of historical social bookkeeping systems. More could be learned of the operation of historical taxation systems and of the means by which the tax collector could be cheated. In this way, the error and bias characteristic of production and wealth statistics might be better diagnosed. We know that in the past, as in the present, it was sometimes to the interest of police and law enforcement agencies to maximize the reported incidence of crime and violence, and at other times minimization of that incidence was to their interest. Consideration of the nature of specific historical situations might allow better assessment of the direction, if not the magnitude, of the bias likely to be characteristic of statistics of crime and violence.

Much more could be done in the way of source, or data, criticism. As an example, two historians using what amounts to an extension of methods of internal criticism have convincingly demonstrated the presence of substantial „biased underenumeration“ in the data provided by the nineteenth century censuses of the United States<sup>11</sup>. Their work calls into question findings reported in a number of studies based on that source and is depressing in that respect. On the other hand, their work, and that of others, also provides grounds for optimism. At a minimum such efforts look toward identification of the categories of research that the censuses will and will not support. Thus these efforts look toward identification of limits. Obviously, if more can be learned of both the direction and magnitude of bias then pitfalls confronted in the use of the source will be reduced or eliminated. And clearly, such efforts can be extended to other categories of source materials.

More effective use of the evidence of the past would require, in other words, larger and more systematic investment of energy and talent in activities that are now often seen — quite wrongly — as preliminary and essentially ancillary to the actual research process. The requirement is not surprising for it has already been encountered in the other social sciences. Sample survey research is a case in point. What might be termed the technology of survey research has become an area of specialization, and a variety of subspecializations in aspects of that technology have appeared which range from sample and instrument design through interviewing

<sup>11</sup> Sharpless, John B., and Shortridge, Ray M., Biased Underenumeration in Census Manuscripts: Methodological Implications, in: *Journal of Urban History*, 1 (August 1975).

techniques to data preparation, management and archiving. If the use of historical evidence in the pursuit of social scientific goals was taken seriously, appearance of analogous specializations and areas of specialized knowledge and activity related to historical source material might be anticipated, encouraged and rewarded.

But these can be only partial remedies for the shortcomings of historical sources. Historical data that were never collected cannot be created; data that was once collected but destroyed cannot be recreated. Even at best, the data of the past will remain incomplete and error ridden and critical variables will be lacking. The capacity of social scientists to use the data of the historical past to develop scientific knowledge of human behavior will remain similarly limited. It may be worth remembering, however, that astronomers also study phenomena that are distant in both time and space. The data of astronomy are in some ways similar to those of history. The astronomer's data is also a residue — the residual radiation from spacially and temporally distant stellar bodies that has penetrated to earth after countless filtering mechanisms have taken their toll. Astronomers have the advantage of the laws of physics which allow diagnosis of the effects of those filters and which facilitate development of compensatory information. A system of social knowledge that equals physic in rigor and precision probably cannot be expected. Even so, an improved social science would increase the utility of historical evidence for the pursuit of scientific knowledge of human affairs, and the study of the past can contribute to that improvement.

The use of quantitative methods and materials by historians has already contributed to a more comprehensive and enriched view of the past. But if historians are to join in the quest for scientific knowledge of human affairs, the agenda of change remains long and crowded, and it leads „quite outside history“. Historians may not choose to abandon the goals of nineteenth century historiography, but even the effort to create a more valid and reliable view of the past requires further and more radical transformation of their craft than has yet occurred.

## Quantitative Analysis of Historical Material as the Basis for a New Cooperation Between History and Sociology

### I. Programmatic Cooperation Versus Quantitative Analysis

There is no shortage of programmatic statements on systematic cooperation between historians and sociologists, but actual joint work has rather been impeded by just such programs. This is not a unique experience, as programs for interdisciplinary work between other disciplines often fared no better<sup>1</sup>. Interdisciplinarity as a sustained activity requires certain conditions quite different from those emphasized in many of the programmatic statements. It is the contention of this paper that these conditions now exist for the quantitative analysis of historical materials.

So far, the most important contributions to sociology were the work of scholars who as individuals were able to synthesize knowledge from sociology and history. Among the several scholars from the founding period of sociology — such as Lorenz v. Stein, Robert v. Mohl, Gustav Schmoller, Werner Sombart, Joseph Schumpeter —, Max Weber stands out as a scholar with a universal knowledge by the standards of his time who translated historical material into a basis for a systematic sociology<sup>2</sup>. Contrary to Weber's reception in the USA and from there subsequently in other countries, his colleagues in Germany saw in him more of a social historian than of a sociologist; von Wiese's reference to Max Weber in his short „History of Sociology“ as a promising empiricist and economic historian is representative<sup>3</sup>. As the knowledge of historical detail accumulates such a synthesis as an individual accomplishment

<sup>1</sup> Examples are the attempts to institutionalize interdisciplinarity between sociology and medicine, jurisprudence, and economics. Cf. Scheuch, Erwin K., *Interdisziplinäre Zusammenarbeit — aus der Sicht des Soziologen*, in: Langenbeck's Archiv der Chirurgie, No. 337, München 1974.

<sup>2</sup> Max Weber as re-imported from the United States and interpreted by Talcott Parsons is primarily the author of part of his incompleted „Wirtschaft und Gesellschaft“. Prior to this „Parsonification“ the work considered central was his sociological analysis of world religions, *Gesammelte Aufsätze zur Religionssoziologie*, 3 vols., Tübingen 1920–1921.

<sup>3</sup> Cf. von Wiese, Leopold, *Soziologie, Geschichte und Hauptprobleme*, 5th edit., Berlin 1954, p. 129 and elsewhere. As is true for many of his contemporaries, von Wiese treated the cultural philosopher Alfred Weber as the more prominent of the brothers.



becomes an obvious impossibility — safe for a selectivity and level of abstraction from details that earn such attempts the epithet „tour de force“. In such uses of universal history as by Herbert Spencer, or Oswald Spengler, or Pitirim Sorokin historical material is not really the object of an analysis but illustration for a systematic point, characteristically some form of either evolutionary or cyclical perspective of human existence<sup>4</sup>.

Actually this was the prevailing use made of history even earlier by some of the founding fathers of sociology, such as Georges Sorel, or Gaetano Mosca, or Vilfredo Pareto — and certainly also Karl Marx, excepting his „18th of Brumaire of Louis Napoleon“<sup>5</sup>. This use of historical material as illustration in the guise of „proof“ contributed to the hostility of historians against sociology, which in the tradition of German historicism was expressed by such influential historians as Johann Gustav Droysen or Heinrich v. Treitschke<sup>6</sup>. In retrospect, it is specifically the use of historical material from a single systematic viewpoint — be it the eternal circulation of elites, or the oscillation between materialistic and idealistic orientation of cultures, or the trend from simple to ever increasing complexity, or history as a succession of class struggles — that makes out of „great books“ very perishable products. As knowledge of historical detail increases these great books suffer the fate that has been characterized for the natural sciences as the greatest tragedy in the life of a scholar: A beautiful idea slain by a brute fact. Courses on the history of sociology have as their main subject matter such systematic uses of historical material by universalistically educated scholars that are now merely of historical interest; and not as contributions of substantive knowledge<sup>7</sup>.

<sup>4</sup> The most important evolutionary writer for sociology has been Spencer, Herbert, *The Principles of Sociology*, 3 vols., New York 1876–1896. A very characteristic recent example of cyclical theories is Sorokin's attempt to interpret history as an oscillation between materialistic and spiritual orientation; Sorokin, Pitirim A., *Society, Culture, and Personality*, New York 1947, especially Part VI.

<sup>5</sup> The evolutionary orientation in nearly all of Marx's works is obvious, although it is not always recognized to which degree Marx chose his references to actual facts and events to fit his evolutionary scheme. In some of his comments on events of his own time, however, Marx is a historiographer — specifically in his analyses of the various uprisings in France.

<sup>6</sup> In reaction to this the school of historicism in Germany emphasized the need to understand each time by itself as a unique configuration. This historical approach had for a considerable time the function of an alternative social science to sociology. Compare Droysen, Johann Gustav, *Grundriß der Historik*, Leipzig 1868 and von Treitschke, Heinrich, *Die Gesellschaftswissenschaft. Ein kritischer Versuch*, Leipzig 1859.

<sup>7</sup> An example of this are theories on the development of the family from a presumed „natural“ condition to its current form. Cases were cited to argue for the primacy of just one form of the family, such as the primacy of group marriage by Friedrich Engels („Vom Ursprung der Familie. . .“), or of the matriarchal family by Johann Jakob Bachofen, or of the patriarchal family as argued by Edward Westermarck. As systematic information about the past and of development these „great books“ are useless.

This tendency to premature high level generalization by many of the founding fathers of sociology aside — although it is still with us and rewarded with reputation —, the synthesis as an individual accomplishment is obviously only fruitful in the early development of a discipline. This has been no different in the cooperation between other disciplines that are rich in material and conceptual apparatus. There is no way around the need for cooperation between scholars from different disciplines who contribute to this cooperation through their distinct competence, such as the specific competence of the historian in judging documents or being able to place them into context, or the skill of sociologists in data analysis. It is more problematic to which degree and especially in which way sociologists and historians can combine their respective problem formulations and conceptual apparatus — a point to which it will be necessary to return.

In view of this, the various programmatic statements for cooperation between the two disciplines are understandable — and yet they have resulted in more damage than good if they were phrased as exclusive programs rather than as one new possibility in addition to other programs (or paradigms, as it now has become fashionable to say). Examples of such exclusive programs are the demand to rewrite history as social history (at least for purposes of instruction in secondary schools), or the blanket demand for history to be practiced as an applied social science<sup>8</sup>. Add to this such ideological formulations as the request that history should from now on spotlight the downtrodden, the victims of events rather than the actors, and the call for cooperation between historians and sociologists becomes a political issue<sup>9</sup>. However, historians may be reminded that this politicization of disciplinary issues is not a consequence of „sociologisation“ but due to a more general trend that produced

<sup>8</sup> Compare Wehler, Hans-Ulrich (ed.), *Geschichte und Soziologie*, Köln 1972; Tilly, Charles, *Clio and Minerva*, in: McKinney, J. C., and Tiryakian, E. A. (eds.), *Theoretical Sociology*, New York 1970, pp. 434–466; Hobsbawm, E. J., *From Social History to the History of Society*, in: *Daedalus*, No. 100 (1971); Benson, L., *Toward the Scientific Study of History*, New York 1972. For problems resulting from such an approach see Sherif, Mustapher and Sherif, Caroline (eds.), *Interdisciplinary Relationships in the Social Sciences*, Chicago 1969. For a more pragmatic approach see the program of the International Association for Historical Social Research, QUANTUM (1975).

<sup>9</sup> A very pointed advocate for a new history whose heroes would be the silent masses, a history that would view events from the bottom up instead of replicating the view of the „makers“ of history, is Modell, John, *Die ‚Neue Sozialgeschichte‘ in Amerika*, in: *Geschichte und Gesellschaft*, Vol.1 (1975), pp. 155 passim. One of the most influential sociologists-plus-historians, Richard Tilly, is not free of the claim that this new history is at last real history. Cf. Tilly, Richard, and Hohorst, Gerd, *Sozialer Protest in Deutschland im 19. Jahrhundert*, in: Jarausch, Konrad (ed.), *Quantifizierung in der Geschichtswissenschaft*, Düsseldorf 1976, pp. 232–278; also Tilly, Richard, *Zum Thema*, in: *Geschichte und Gesellschaft*, 3 (1977), pp. 151–152. The ideological use of history has invaded the class rooms of secondary education and journalism. Cf. Rudolph, Hermann, *Was ist Geschichte?*, in: *Frankfurter Allgemeine Zeitung*, June 9, 1978, p. 25.

also such sects as „revisionistic“ history<sup>10</sup>. There is supreme irony in this reideologization of history and sociology alike, as it is largely based on a defunct historiography.

The currently virulent ideologies apart, programmatic requests for cooperation of the type quoted above tend to block sustained work for reasons of principle. In all these cases a follow-up of the programmatic request assumes that from now on a particular perspective, a paradigm, is shared, up to and often including a common teleology. In this day and age this may be the binding element for a sect but cannot be the universal orientation for an empirical discipline that every so often happens upon new knowledge.

The quantitative analysis of historical material may sometimes be advocated in a manner that sounds like the programs referred to earlier, and yet it is a completely different basis for cooperation. All that is required here is an agreement on a common material, and a common technology in data handling. From traditional points of view in the discipline this may presuppose both substantive and methodological decisions that are considered alien. One may object that trivial objects of trivial people are not the observational base for a history that reveals purposes to mankind or is able to provide lessons to the present. As a sociologist, these arguments are outside my proper realm of interest. The methodological argument that this quantitative history presupposes a deterministic view, however, is not; this was the central issue in one of the great methodological controversies in sociology, namely the historicism controversy. As is usual for such sweeping issues it did not get resolved but was largely forgotten — and rightly so. In order to perform quantitative analysis a deterministic view is unnecessary; it is only necessary to expect that there are also regularities in human existence which are not apparent to the actors themselves but have to be inferred. Whether this is indeed so and what strength these factors have relative to unique influences, is an empirical question. The experience so far suggests that it is worthwhile to continue this search, and be it only as a form of description that transcends any observers ability. And if some sociologists argue that with industrialization man's conditions are changed in a way that reference to previous experience is an obstacle to what is really needed, namely utopian phantasy, then again this is properly an empirical issue and not a decision immune to it. Anyway, so far the predictions of the non-utopians have been better than utopian scenarios of the immediate future.

<sup>10</sup> A central figure for the evolution of an ideologically committed history is Moore, Barrington, *Social Origins of Dictatorship and Democracy*, Boston 1966. Compare also Rothman, Stanley, Barrington Moore and the *Dialectics of Revolution*, in: *American Political Science Review*, Vol. 64 (1970), pp. 61–82, and Fogel, R. W., and Engerman, S. L. (eds.), *The Reinterpretation of American Economic History*, New York 1971. French „structuralism“ is another intellectual fashion that encouraged an ideologically committed history; compare Schiwy, Günter (ed.), *Der französische Strukturalismus*, Reinbek 1969. A German variant of this plea for ideological commitment is Schmidt, Alfred, *Geschichte und Struktur*, München 1971.

Quantitative analysis of historical material provides a common empirical base for many diverse interpretations — in this way similar to such a tool as time and money budgets<sup>11</sup>. It's particular contribution is the description of diversity and the detection of regularities in so far as both transcend the observational powers of contemporaries — and this is a direct analogy to the most fruitful applications of quantitative techniques in sociology. This empirical base is open to a variety of paradigms — and emphasizing this may help to overcome some of the reservations of historians that as yet view this trend with reservations.

For sociologists a different explanation is necessary to stimulate their attention. There are, however, two traditions that impede the full use of a new, vast empirical base for their discipline.

## II. On the Evolutionary Tradition in Sociology

The topic should be unnecessary as we have it on the eminent authority of Talcott Parsons that „Herbert Spencer is dead!“<sup>12</sup> However, Herbert Spencer under different names is very much alive, indeed, and kicking for the same reasons that produced Herbert Spencers in the first place. For a while it seemed that Herbert Spencer was dead, as the motivations for the evolutionary canvasses in our disciplines had paled. Now, the interest in the course of development is high once again as the confidence in the acceptability of the future is low.

Sociology — in the form that has become professionalized — is indeed a „crisis discipline“. In the 19th century there was a wide-spread agreement that the current situation, the current societal condition, could not last. This was not to be a new form of human existence to continue but a transitory period<sup>13</sup>. Conservative observers, such as Wilhelm Heinrich v. Riehl or Lorenz v. Stein, might emphasize the features of dissolution that they saw at their time, and would accordingly choose

<sup>11</sup> This was the perspective from which Lenin advocated the collection of time and money budgets, namely as reflections of reality. This led to a specific version of empirical research in countries that officially follow Leninist principles. The most representative presentation of this research is Szalai, Alexander (ed.), *The Use of Time*, The Hague 1972.

<sup>12</sup> Cf. Parsons, Talcott, *The Structure of Social Action*, New York 1937. Later, Parsons is less certain, as is evident in the *Point of View of the Author*, in: Black, Max (ed.), *The Social Theories of Talcott Parsons*, Englewood Cliffs 1961, pp. 311–363.

<sup>13</sup> That in spite of all the protestations about his „scientism“ in charting the course of history, Karl Marx is really driven by apocalyptic vision was recognized early by Sorel, George, *La décomposition du marxisme*, Paris 1907. René König revived this understanding in *Soziologie heute*, Zürich 1949, pp. 30 *passim*, and he stimulated the work of Jakob Taubes, *Abendländische Eschatologie*, Bern 1947.

topics and perspective in empirical work. From a more radical perspective one might emphasize the direction of development and opt for a teleology, which is obvious in the works of Marx, Spencer, Comte, but also characteristic in such concepts as Ferdinand Tönnies' „Gemeinschaft“ and „Gesellschaft“<sup>14</sup>.

For sociology, the empirical basis for the construction of these teleological schemes, answering „whither are we going?“, shifted over time with material from other disciplines being dominant then. We are now used to sociologists being their own data gatherers but during this ‚heroic‘ period they relied on historical material, sometimes ethnographic material. There was a preference for historical material until about the 1870s, and as subsequently ambitious ethnographic reports became available this was the preferred material. The differences between Marx and the elder Friedrich Engels are a case in point<sup>15</sup>. To a degree both types of material were used in the same way: one would look at the past or at „primitive“ cultures as a description of origins. Hopefully, one would find a few examples of intervening conditions, and from there constructed a picture of the future.

A specific interest was the search for zero-points of human developments, those elementary forms behind which human existence did not go back, the bases from where human existence progressed. And if one had found the zero-base, one could then speculate to which degree human history was in error, impeded possibilities; not only would one be able to predict the future but to create a better one by knowing from history not only the direction but also man's unused potential. Seen from today it may be baffling to read the arguments about the originality either of the nuclear family or of group-marriages, the arguments resting on exceedingly few cases. However, this was primarily not a discussion with scholarly intentions, *sine ira et studio*, but one which had immediate ideological consequences. The writer to whom we are obliged for the very term „sociology“, Auguste Comte, consequently proceeded from an apparent concern with scholarship to the founding of an elitist sect.

<sup>14</sup> The apparent dichotomies prevalent in sociology early in this century, were frequently teleological in an extremely reduced form. This is true for Emile Durkheim's juxtaposition of mechanical vs. organic solidarity, for Ferdinand Tönnies' dichotomy „Gemeinschaft“ vs. „Gesellschaft“, and for Charles Cooley's distinction primary vs. secondary groups. The very basic concepts of sociology until the recent past implied teleologies.

<sup>15</sup> Friedrich Engels was the consumer of ethnographic material, as he was attentive to empirical material that came into his view. However, there are significant blind spots, the most important being an ignorance of the quantitative history already available at that time, such as Graunt, John, Natural and Political Observations Mentioned in a Following Index and Made Upon the Bills of Mortality, London 1662. Neither did they pay attention to quantitative research at their time, such as Le Play, M. F., *La reforme sociale en France*, Paris 1864, and Morselli, Henry, *Suicide – An Essay on Comparative Moral Statistics*, New York 1882 (Durkheim's „Suicide“ was not published until 1897!). In addition there were many more good statistical sources available than were used – as is evident from Weiss, Hilda P., *Les enquêtes ouvrières en France entre 1830 et 1848*, Paris 1936. The social sciences – or at least the part that was handed-on to the past as important – could have been far more empirical than they actually were.

Of course, the empirical base available to social scientists of this time was extremely thin. Each time when a significant new contribution of ethnographers became available the evolutionary constructs needed to be rearranged. Equally, a single case, the presumed conditions in an individual tribe, had a sensational impact provided the case fitted the preconceptions of the social scientist. An example is the publication by Henry Morgan about the Iroquois. Morgan was employed as an engineer in building a railroad to Lake Erie, and he became fascinated by the life of Indians as he was able to record it at the end of the 19th century; there are now some arguments that this was a non-typical situation for the tribe itself. Even though this was a contribution by an amateur, it was immediately used by the evolutionists such as Friedrich Engels, and even today the presumed case of the Iroquois as proof of the primacy of matriarchalism was cited uncritically by ideologists such as Ernest Bornemann.

The empirical material was in truth not an empirical base for the theory but mere illustration for preconceptions. Thus, when the cultural revolution of the sixties erupted with the dusting off of 19th century thought, the example of the earlier use of the Iroquois had a contemporary parallel. Some deservedly forgotten student-sociologist thought he had found an African tribe, the Amba, who lacked any stratification in power or authority. And significantly, a fully grown German university professor, Ralf Dahrendorf, argued the case as though it would be decisive for the question whether stratification in power is a necessary part of a developed social structure<sup>16</sup>. Contrary to the situation at the time of Henry Morgan and his report on the Iroquois, there was now ample ethnological material on the stratification in power of tribal societies, but the neo-evolutionists were not interested in this. Even if the Amba had indeed lacked any stratification in power and authority: So what?

The use of historical material by evolutionaries and neo-evolutionaries was of the same character. This was not really an interest in history as a characterization in each case of past conditions as they really were, and there was accordingly no immersion in sources. The characteristic evolutionist was and is in search of building blocks to fit his architectural design of human development. Contrasting the use of ethnographic and historical material by Johann Jakob Bachofen and Edward Westermarck with the work of Karl Wittfogel or Max Weber exemplifies the difference between using other disciplines for illustration rather than as providing an extension of the empirical base for sociology.

When Bachofen and Westermarck argued for one „original“ form of the family, then „original“ was to imply „natural“. Human history was then a formation of this natural state as a deformation, until it would be possible now to regain the natural state at a higher level of civilization. The mystical theologian Bachofen cited historical and ethnographic material, but also used legends and fairy tales, to demonstrate the primacy of matriarchalism; the historically known forms of the family, such as the classical Roman family (or rather what at that time was believed to have

<sup>16</sup> Cf. Dahrendorf, Ralf, Amba, Amerikaner und Kommunisten — zur These der Universalität von Herrschaft, in: Dahrendorf, R. (ed.), *Pfade aus Utopia*, München 1967, pp. 315–336.

been the Roman family) or the family of Judaism were seen as suppression of a natural state<sup>17</sup>. Parallels were maintained between the presumed suppression of women in patriarchalism and the political organisation of countries. Westermarck, too attributed a paradigmatic quality to the forms of the familial distribution of authority, although he cited historical sources for the primacy of the patriarchal family. Even though this controversy surfaced again as part of the intellectual imitations that were characteristic for the „cultural revolution“, family sociologists generally agree that searching ethnographic and historical records for a „natural“ state of human existence is futile. Evidence from research on primates makes it more likely that there was „originally“ more than one form of the family. It is characteristic for sociological evolutionism that it is clandestinely anti-historical, namely the search for non-historical conditions and the perspective of history as deformation.

In analyzing historical material on the great river-valley civilizations, Karl Wittfogel also had an ideological motivation, namely to develop a scheme for the necessary development of state socialism into a bureaucratic oligarchy<sup>18</sup>. The centrality of the single source of wealth, the river water, and the need for regulation of this resource, leads to the development of a central bureaucracy — and according to Wittfogel it does so with inevitability. Wittfogel's writings resemble classical evolutionism in his use of history in so far that historical instances of what Wittfogel calls „hydraulic civilizations“ are presented with the intent to demonstrate an inevitable development, in this case the dominance of a bureaucratic class. However, Wittfogel does attempt to work as an historian, and above all history is treated as a normal state of human existence instead of a transitory condition.

<sup>17</sup> Our understanding of the Roman family is largely a reflection of the construction of ideal types by legal historians. Even if we leave aside the question whether these legal constructs had much to do with reality — and among other indications sculptures and inscriptions on cemeteries suggest otherwise —, there were two legal forms for marriage among which the spouses could choose. The patriarchal family was the marriage „cum manu“, the essence of which was the transfer of the wife from her kin to that of her husband's, as against the marriage *sine manu* which was a contract between individuals including the right to divorce for both parties. This latter form was usual and disapproved by the Caesars — which may be the reason for historians to be largely silent about it. Even the family of Ancient Judaism was probably not an institution of despotic power as it appeared in official descriptions and in several spectacular cases in the Old Testament. At the Institute of Applied Social Research of the University of Cologne we reanalyzed the conflicts within the family that are described in the Old Testament. These descriptions were read as an indication of what they implied about the operative norms in daily life. As a result we concluded that the usual picture of Patriarchalism in Ancient Judaism referred primarily to the family as a religious unit, and in official transactions with the outside world — but not in other fields of behavior. Cf. Wurmnest, Karl Friedrich, *Die Rolle des Individuums innerhalb von Familie und Ehe im Alten Testament anhand relevanter Texte unter Berücksichtigung welt- und kulturgeschichtlich bedingter Raum-Zeitstrukturen*, unpublished dissertation at the Philosophische Fakultät, University of Cologne, 1978.

<sup>18</sup> Cf. Wittfogel, Karl A., *Oriental Despotism*, New York 1957. Wittfogel presents his monograph as comparative research with total societies as a unit.

We mentioned already that to his German contemporaries, Max Weber was rather an economic historian than the theorist of the first part of *Economy and Society*<sup>19</sup>. At the beginning of his career, German economic historians were analyzing their material in order to show necessary „stages“ in the development of civilizations, and to demonstrate a close relationship between an economic and a social order. This was a far cry from the evolutionism of Auguste Comte or Herbert Spencer who maintained a continuity of evolution from simple inanimated conditions to the complexity of society, an evolution that presumably was inevitable, monodirectional and monocausal. Yet Weber differed from those contemporary economic historians still further into the direction of an historian *strictu sensu*. In the central part of his work, the volumes on the sociology of religion, Weber deliberately varies civilizations in order to refute monocausal notions about the relation between „base“ (economy) and „superstructure“ (religion): Each of these civilizations has to be understood via its own „Sinn“ (approximately „meaning“), has a „Gestalt“ (approximately „shape“) of its own<sup>20</sup>. Yet Weber was also a sociologist using concepts without specific time-space meanings, and in this context he was a modified evolutionist. Weber's writings on music, on authority (an unfortunate translation of his „Herrschaft“), on science, and on bureaucracy all have one „Leitmotiv“: Why did a specific type of rationality develop only in Europe?<sup>21</sup> In pursuing these two main lines of work — their relation cannot be discussed here — Weber did not work with the conclusions of historians but with the source material itself.

<sup>19</sup> In the American reception of Max Weber the conditions under which Weber approached his monumental *Economy and Society* are largely forgotten. It is no longer possible to reconstruct a definitive version of this posthumous work, as it is likely that Weber changed his original notions several times as the work progressed. *Economy and Society* was to be in a way a contrast to his work so far as it was to present his concepts in a systematic way. This proved to be more difficult than expected as indeed the concepts were developed at different times in response to different tasks. Thus, it is simply not possible to establish a systematic relation between the taxonomies for forms of legitimate authority („reine Typen der Herrschaft“) and the taxonomy for types of action orientation („Typen des Handelns“), without creating confusion — as is indeed sometimes the case in *Economy and Society*. Cf. Scheuch, Erwin K., and Kutsch, Thomas, *Grundbegriffe der Soziologie*, 2nd edit., Stuttgart 1975, Chapter 9, Sections 1 and 2. For a new way to look at *Economy and Society* see Tenbruck, Friedrich H., *Abschied von Wirtschaft und Gesellschaft*, in: *Staatswissenschaft*, Tübingen 1978, pp. 1–34.

<sup>20</sup> The notion of „Sinn“ is central for Weber's analysis of a total system: its the attribution of „Leitmotivs“ to the actions in a society. Cf. Girndt, Helmut, *Das soziale Handeln als Grundkategorie erfahrungswissenschaftlicher Soziologie*, Tübingen 1967.

<sup>21</sup> The notion of „rationality“ as a „Leitmotiv“ of systems is explored in Münch, Richard, *Max Webers „Anatomie des okzidentalen Rationalismus“ — eine systemtheoretische Lektüre*, in: *Soziale Welt*, 29 (1978), pp. 217–246. There are two more sides to Weber's work, the second of which is largely unknown today. It is better known that Weber was interested in methodological issues, as is evident e. g. in *Weber, Max, Methodologische Schriften*, Frankfurt 1968 (a collection), but he was also a passionate commentator on political developments. Cf. *Weber, Max, Gesammelte politische Schriften*, 2nd edit., Tübingen 1958.



Even the second accent of Weber's work could not be replicated today. There may be a revival of evolutionism in intellectual life, but only in the sense of a philosophical exploitation of historical generalizations and not in the sense of a use of historical sources to construct laws of development. None of the grand conclusions of the evolutionists stood the test of time, and it is unlikely that the neo-evolutionists will fare better. There is now such a wealth of empirical evidence, and the movement of quantitative history increases the volume still further, that a simple ordering whether in „stages of development“ or in cycles is no longer feasible. The publications of Shmuel Eisenstadt demonstrate that historiography and social science can still be combined in the grand style, but the accent is on comparativism and definitely not on evolution<sup>22</sup>. A sociology that hopes to regain the courage to sweeping theories of the 19th century, a sociology that looks upon history as an opportunity to revive evolutionism, misses the specific usefulness of the current meeting of sociology and history. The description of everyday life and mass events in the past that now becomes possible, definitely does not lend itself to a type of theoretizing in the evolutionary tradition. Although at first sight the assertion may seem paradoxical, it nevertheless can be argued that structural-functional theoretizing is more compatible with the data from quantitative history.

### III. Is Functionalism Necessary Anti-Historical?

This is only in part a rhetorical question, as there is no unequivocal answer: there is no necessary conflict between a structural-functional kind of theoretizing and history, but in practice this is so. This is probably due to the development of structural-functionalism in the United States. Be that as it may, structural-functionalism has been so dominant a mode of theoretizing since the middle forties until the middle sixties that it became synonymous with general sociological theory. In practice, this kind of theory prided itself in formulating general sentences without time-space referents, was general theory in line with the introductory part of Weber's *Economy and Society* and not with his other writing. It was usual amongst sociologists to understand this mode of formulating as following the example of the successful natural sciences — and that meant largely physics.

This is, however, a misunderstanding. The discipline in the natural sciences closest to structural-functionalism in sociology is, at least in the case of Parsons, rather biology. Biology, that is a discipline which has to do with reactive systems, and in

<sup>22</sup> A good introduction to his approach is Eisenstadt, S. N. (ed.), *The Decline of Empires*, Englewood Cliffs 1967. Much more ambitious is Eisenstadt, S. N. (ed.), *Political Sociology*, New York 1971, which in spite of its title is predominantly a book with historical comparisons, albeit of a non-quantitative character.

this sense it is contrary to some of the classical sciences. Here, the object is not „cause“ and „effect“ but „effect and counter-effect“. At any given time, an object or process may serve more than one function, or the same object or process may serve different functions at different times. A biological organism as an object of explanation is a vastly more complicated thing than the inanimate nature. Society as well, if structural-functionalism is properly practiced, is treated as a reactive system and not in an analogy to inanimate nature.

This would be complicated enough, but in addition there is an unnecessary problem in the functionalism as it is actually practiced. It becomes most apparent in what is called „Systems Analysis“. In this approach it is assumed or implied that basically all parts of a system are necessarily cooperating and that they react tightly together. This is completely unnecessary to assume since there are parts in the body too, which are unnecessary, not everything is directed to the same purpose. There are countervailing processes, functional substitutes in addition to fixed organs, and a lot of give-and-take, i. e. looseness between organs and parts of a body. However, systems theory as a specific form of structural functionalism in its actual practice assumed a direct reaction of all parts of a system to each other<sup>23</sup>.

The conceptual apparatus, the research problems and the empirical research connected with these approaches found its purest expression in small-group research<sup>24</sup>. Indeed, small-group research has as an object something that does not really exist but is constituted as an construct — and yet this research was to a degree successful in finding universals that eluded sociologists in many other areas. Yes, sociology has developed universal sentences about human behavior that can be applied in a variety of contexts. This copy of physics was not a story of complete failure, unfortunately it is also not a story of a large scale success. As sociologists moved beyond the micro level it became much harder to justify time-space free sentences in terms of „X“ being a function of „Y“. What stood sociologists in good stead, namely the type of conceptual apparatus, the type of methodology and specifically the type of interpretation when they worked with the immediately observable, was much less successful when they had to work with indicators and the proof had to be inferential. Most macro phenomena are of an inferential nature. This became even more important and more obvious when cross-cultural research became important.

<sup>23</sup> The most prominent representative of this kind of systems theory in Germany is today Niklas Luhmann; cf. Luhmann, Niklas (ed.), *Soziologische Aufklärung — Aufsätze zur Theorie sozialer Systeme*, Opladen 1970; also *Zweckbegriff und Systemrationalität*, Tübingen 1968; also *Zur systemtheoretischen Konstruktion von Evolution*, in: Lepsius, Rainer (ed.), *Zwischenbilanz der Soziologie*, Stuttgart 1976, pp. 37–48; also *Generalized Media and the Problem of Contingency*, in: Loubser, Jan J., et al. (eds.), *Explorations in General Theory in the Social Sciences*, New York 1974.

<sup>24</sup> A recent overview of the whole field is Schneider, H. D., *Kleingruppenforschung*, Stuttgart 1975. For the self understanding of this approach see Bales, Robert Fred, *Personality and Interpersonal Behavior*, New York 1970. The artificial character of the whole field is critized by Sorokin, Pitirim A., *Fads and Foibles in Modern Sociology and Related Sciences*, Chicago 1956.

As structural functionalism has been a part-success, as there are areas which can be shown as models to other disciplines, as the methodology works very fine, this partial success tends to somewhat impede the openness in turning to such a vast new area of material as becomes available to us in quantitative history. Especially, the part-success tends to inhibit a re-examination whether structural-functionalism needs to be practiced in the way that prevailed up to now.

#### IV. Empirical Sociology Experiences Limits

At the end of a period of more than thirty years of development in empirical sociology, there is now some soul searching and attempts at stock taking<sup>25</sup>. This was in many ways a most successful period: in some fields general „laws“ akin to those of physics were identified; the methodology for the social sciences in general was furthered and became an export article even to those who voiced programmatic reservations against „positivistic“ sociology; and a vast amount of descriptive knowledge was accumulated. Methodology and social description could be so standardised that they could be the base for a service industry that now produces vast quantities of facts. Increasingly, social scientists begin to tap the additional vast data resources that come into being as a side-product of public and private bureaucracies<sup>26</sup>. Now that we are relatively data rich, we begin to feel just as those rich in other proper-

<sup>25</sup> Examples of this self doubt, coupled with the desire to retain the claim to be at the same time a science and a tool of the Enlightenment are Birnbaum, Norman, *The Crisis of Industrial Society*, London 1969; Gouldner, Alvin W., *The Coming Crisis of Western Sociology*, New York 1970; Dahrendorf, Ralf, *Die Soziologie und der Soziologe*, in: Hess, Gerhard (ed.), *Konstanzer Universitätsreden*, no year. It is instructive to compare these diagnoses with the actual work presented at the 17th German „Soziologentag“ in 1974, presumably a crisis year if judged by public appearances: During the sociological convention routinized science (in the sense of Th. Kuhn) prevailed. See Lepsius, *Zwischenbilanz*. See also Scheuch, Erwin K., *Die wechselnde Datenbasis der Soziologie. Zur Interaktion zwischen Theorie und Empirie*, in: Müller, Paul J. (ed.), *Die Analyse prozeß-produzierter Daten*, Stuttgart 1977, pp. 5–41.

<sup>26</sup> An over-all view of this vast area is Wilcox, Lesly D., et al. (eds.), *Social Indicators and Societal Monitoring*, Amsterdam 1972. So far social scientists use only fractions of the material existing, as can be inferred from Statistisches Bundesamt, *Das Arbeitsgebiet der Bundesstatistik* 1976, Stuttgart 1976. Currently, the chief interest in using these process-produced or officially collected data is their appropriateness for societal monitoring, as explained in Zapf, Wolfgang (ed.), *Sozialberichterstattung – Möglichkeiten und Probleme*, Göttingen 1976. See also Krupp, Hans-Jürgen, and Zapf, Wolfgang, *Sozialpolitik und Sozialberichterstattung*, Frankfurt 1977, for a characterization of the most important research unit in this field in Germany, SPES: Zapf, Wolfgang (ed.), *Soziale Indikatoren*, 3 vols., Frankfurt 1974–1975. This field is internationalizing fast, as can be derived from international data collections such as EUROSTAT: Social

ties presumably do: it is great to be rich but it satisfies a lot less than expected. Many of the facts and figures are suspected to be less informative than we thought at a time when each new fact or figure possessed a novelty value.

To give one example of considerable personal importance. During the fifties it was empirically demonstrated again and again that one large difference between mass opinion in Europe and in the United States was what political scientists conceptualized as „system trust“. Europeans were shown to be highly sceptical about their politicians, their political parties, and sometimes also of all of the political system. In contrast, respondents in the United States expressed an unshakable respect for the office of the president and the institution of the two-party-system, even when they detested a particular president or found their two political parties at a given time to be in terrible shape. Just as they were reputedly cynical about morals, these Europeans were called political cynics, and American political scientists concluded that this was not a condition in which a meaningful democracy could flourish. Now that we count the year five post-Watergate the trust of Americans in their political institutions is below that which opinion researchers now report for European countries<sup>27</sup>. What did we measure some thirty years ago: Was it really an aspect of a distinct political structure, or merely a mood? And do changes in mood matter very much in the operation of a political system?<sup>28</sup>

In looking back at over thirty years of data collection we can observe both high stability of differences between countries and groups within a country for some subject matter, and great changes up to fickleness of figures in other areas. In the field of leisure we have witnessed a high instability of behavior, and this is currently especially true in research on tourism<sup>29</sup>. Research on sexual matters has shown a tremendous instability in beliefs and opinions, and far more stability than instability in behavior. Currently, there is in Germany a debate whether we witness a major change in values amongst youth — the school of „post industrialism“ believes that this is so —, and whether the traditional work ethic is falling apart; it is by no means clear what the figures really do indicate. Where do we measure a structural property, where do we record a mere transitional state? Sociologists are becoming — albeit a bit too slow — more careful in interpreting numbers.

Indicators for the European Community, Luxemburg 1977; and there is also now an international newsheet: Social Indicators Newsletter, Social Science Research Council, New York. While much of this work is pure induction, there are attempts to develop a rationale as in Fox, Karl A., Social Indicators and Social Theory, New York 1974; OECD, Measuring Social Well-Being, Paris 1976. Decisive for the expansion of basic research using these resources will be the development of an appropriate infrastructure of data services, as reported by Rokkan, Stein, Data Services in Western Europe — Reflections on Variations in the Conditions of Academic Institution-Building, in: American Behavioral Scientist, 19 (1976), pp. 443–454.

<sup>27</sup> Cf. Huntington, Samuel, et al., The Crisis of Western Democracy, New York 1976.

<sup>28</sup> Critical of the literature on the loss of governability is Scheuch, Erwin K., Wird die Bundesrepublik unregierbar?, AGV Metall, Köln 1976.

<sup>29</sup> See Scheuch, Erwin K., and Scherhorn, Gerhard, Soziologie der Freizeit und des Konsums, in: König, Rene (ed.), Handbuch der empirischen Sozialforschung, Vol. 11, Stuttgart 1977.

Empiricism was quite successful in providing a basis for micro sociology. Macro sociology, however, did not progress in the way it was hoped. In Germany it was especially a group of sociologists sometimes called by others the „Cologne School“ that had advocated cross-level analysis and corresponding data collection as the methodology appropriate for macro sociology<sup>30</sup>. However, the pay-off of this theoretically sound notion has been far less than hoped for<sup>31</sup>. Whether this is due to the empirical research, or the conceptualization of it, or the far greater complexity of an empirically founded macro sociology is an open question.

Research tools are proliferating at a very rapid rate. Techniques that were known for a long time but little used, such as complicated sampling techniques for subgroups of large populations, or techniques of content analysis – are now being actually used. The machinery of large scale electronic data processing is important in turning esoteric knowledge into practical procedures. There are many original ideas in developing so-called unobtrusive techniques i. e. highly inferential measures independent of verbal statements<sup>32</sup>. And in general, there is a greater willingness to combine measurements from several sources: Sociologists may become as critical of their data as historians reputedly are of their sources. On the other hand with the explosive growth of analysis opportunities there has been a tendency to overanalyze some data. The debate about weak versus strong measurement indicates that there has been an unthinking preference for the most powerful statistical techniques regardless of the level of measurement and the reliability of a figure<sup>33</sup>. The latter is a tendency that quantitative historians should better watch.

<sup>30</sup> An overview is given in H. J. Hummel, an adherent of the „Cologne School“, *Probleme der Mehrebenen Analyse*, Stuttgart 1972. A very optimistic expectation was formulated during the sixteenth German „Soziologentag“ in Frankfurt 1968 by Scheuch, Erwin K., *Methodische Probleme gesamtgesellschaftlicher Analysen*, in: Adorno, Theodor W. (ed.), *Spätkapitalismus oder Industriegesellschaft?*, Stuttgart 1969, pp. 153–182.

<sup>31</sup> Massive secondary analysis of data on voting behavior by Franz U. Pappi found in the end that including contextual variables added very little to the explanatory power of the routine individual variables; cf. Pappi, Franz Urban, *Sozialstruktur und politische Konflikte in der Bundesrepublik. Individual- und Kontextanalysen der Wahlenentscheidung* (in publication).

<sup>32</sup> The „classical“ source on unobtrusive techniques is Webb, Eugene, et al., *Unobtrusive Measures*, Chicago 1966. A prerequisite for the large scale use of quantitative content analysis is their combination with sampling techniques; see Kops, Manfred, *Auswahlverfahren in der Inhaltsanalyse*, Meisenheim a. G. 1977. Important contributions to methodology that are especially useful for quantifying and analyzing historical material are Steinhausen, Detlef, and Langer, Klaus, *Clusteranalyse*, Berlin 1977, and Sodeur, Wolfgang, *Empirische Verfahren zur Klassifikation*, Stuttgart 1974. An overview of research techniques that includes advanced methods relevant to quantitative history yet accessible to the non-specialist in methodology is van Koolwijk, Jürgen, and Wicken-Mayser, Maria (eds.), *Techniken der empirischen Sozialforschung*, vols. 2–7, München 1974–1977.

<sup>33</sup> For the debate on the level of measurement appropriate to the data see Scheuch, Erwin K., *Forschungstechniken als Teil der Soziologie heute*, in: Lepsius, Zwischenbilanz, especially pp. 94 *passim*. Compare also Acock, Alan C., and Martin, David, *The Undermeasurement Controversy*, in: *Sociology and Social Research*, 58 (1974), pp. 427 *passim*.

There is now some better understanding of what John Stuart Mill meant when he argued that the social sciences were „observational“, and why Emile Durkheim was anti-experimental. One does not have to reject the experiment as a tool of research in order to sympathize with the notions about the character of social systems that lead to the anti-experimentalism of Mills and Durkheim. Many social phenomena have meaning depending on contexts, are interconnected and multifunctional. Even elementary activities such as eating or sexual intercourse carry several meanings! Social processes are both over- and underdetermined. Relating single variables to each other does usually not do justice to the structure of social phenomena, and with the realization of this condition, analysis techniques are being developed that are more appropriate to the interconnectedness and multidimensionality of social phenomena. Path analysis, causal analysis and LISREL are examples for this trend<sup>34</sup>.

It is doubtful that the limits in explanation which empirical sociologists now sometimes encounter can be overcome solely by further analysis techniques, and a more systematic combination of data. For many problems longer periods of observation are required, and an extension of conditions under which behavior is observed. Quantitative history can provide this extension of the data base for sociology — not so much in quantity but more importantly in quality. In turn, the response of sociology to the multicollineality of relations between variables, the reaction to the multidimensionality of social phenomena, means that today sociologists can offer much more adequate techniques of data handling than would have been possible only ten years ago.

The recent meeting of historians with sociologists, in Germany connected with such names as Hans-Ulrich Wehler and Hans Mommsen, has not necessarily been the most helpful experience<sup>35</sup>. These historians hoped to borrow concepts and generalizations from sociology to regain a larger scope for the discipline of history that appeared to be bogged down into historiographic details. This was an inopportune time to do so, leaving aside the question whether there was ever an opportune time for this.

It was a time when many of us realized that our concepts were more time-and-space-bound than we had so far suspected. For some sociologists it was also a time for a „paradigm change“ — away from systems analysis with its harmonistic view of biology. We now understand biology in a very different way, namely as the discipline of imperfectly constructed beings, as of organisms that side-by-side are characterized by surpluses and deficiencies. Real social systems are evidently imperfectly integra-

<sup>34</sup> Cf. Ziegler, Rolf, *Theorie und Modell*, München 1972; also Blalock, Hubert M., *Causal Inferences in Nonexperimental Research*, Chapel Hill 1964; also Weede, Erich, *Hypothesen, Gleichungen und Daten*, Kronberg Ts. 1977.

<sup>35</sup> As a source for this approach that could be called the sociologization of history instead of the shared use of historical data, consult Wehler, *Geschichte und Soziologie*. By now there are side-by-side several forms of cooperation between sociology and history, as is evident from the range of contributions in Ludz, Peter Christian (ed.), *Soziologie und Sozialgeschichte*, Special issue No. 16, *Kölner Zeitschrift für Soziologie und Sozialpsychologie* (1972) — especially the introduction by the editor.

ted, and by now it is no longer very easy to say what the boundary of the systems is that we are analyzing. We cannot simply use national boundaries as being also system boundaries, as the nation state is coming apart as the highest level of integration. Devolution within nation states and international connectedness make the nation state level just one of several levels that indicated system boundaries. This is an intellectually richer and more flexible sociology, but it is certainly not one from which one could easily borrow ready concepts and generalizations.

## V. Quantitative Analysis of Historical Material as an Extension of Comparativism

A more fruitful orientation in seeking a cooperation between sociology and history is cooperation in exploiting a new data base. Time budget research offers an example for the character of such a research. The use of time is a social indicator lending itself to several interpretations, an indicator that can be put to many uses<sup>36</sup>. In some of the socialist countries, time budget data are employed for such engineering purposes as the calculation of waste times, while the very same data are used by Western social scientists to identify the networks of daily intercourse. Many of the data of quantitative history have the same indeterminate character as time budgets have. Viewed methodologically, most analyses of quantitative history have the character rather of secondary analysis than of primary analysis<sup>37</sup>. This may often cause problems in interpretation, but it does also facilitate cooperation between scholars from different disciplines and with different approaches: They do not need to agree on problem formulations, or concepts. Thus, in looking at quantitative history as an opportunity for secondary analyses of vast quantities of data about previously inaccessible topics and subjects, the pitfalls of the above mentioned approach — the Wehler-Mommsen problem — is avoided.

It is dangerous when sociologists by themselves quantify and analyze historical data, as they usually lack the familiarity with the contexts of these data; and it is no less hazardous if historians feel confident to order high powered statistics from the now easy-to-use packages. But cooperation between sociologists and historians properly goes beyond such a symbiosis in research technology. Historians rightfully expect that quantitative history will give new impetus to history as a generalizing

<sup>36</sup> Cf. Szalai, *The Use of Time*.

<sup>37</sup> „Secondary“ does not imply „second class“ but denotes a use of data different from the intentions with which the data were collected. The classical source on the methodological issues in secondary analysis is Hyman, Herbert, *Secondary Analysis of Sample Surveys — Principles, Procedures, and Potentialities*, New York 1972.

discipline, and sociologists hope for a vast extension of their empirical base. In this latter sense the use of historical data is a form of comparativism, is observation under varying conditions in the sense that John Stuart Mills argued for „observational“ social sciences. This form of comparativism complements and extends significantly what currently is being done in comparative social research.

One of the important resources for sociological comparativism has always been — earlier more so than during the last decades — ethnology, and here a development analogous to that now in quantitative history occurred much earlier. A group of ethnologists around John Peter Murdock from Yale translated the ethnographic reports of their time into a common scheme<sup>38</sup>. This meant among other things that check-lists had to be developed for institutions and fields of behavior as a prerequisite for the coding of ethnographic descriptions. Methodologically, this implied the translation of descriptive accounts into configurations of variables. Only through this „translation“ becomes it possible to develop a quantitative ethnology on a world scale as though the descriptive accounts had been questionnaires about cultures: Frequencies are identified, correlations are computed, factor analyses are meant to show hidden communalities. By now the „Human Relations Area File“ (HRAF) is in part machine readable, and available in several countries. While this increases its accessibility, and makes comparative ethnography something every graduate student can practice, the decisive step was not the machinery but the „translation“ of the narratives. The organization of data sets from projects in quantitative history could do the same for historical data<sup>39</sup>.

It would be of considerable greater consequence. The Human Relations Area File has data from more than 500 cultures, and while there are greater variations between, the hundreds of simpler cultures, they remain simple cultures that are of limited relevance for the understanding of a complex modern society. Even though the volume of quantitative history has been limited, at least as compared to quantitative ethnology, its impact for social science has been far greater<sup>40</sup>. The conditions

<sup>38</sup> Cf. Murdock, George Peter, *Social Structure*, New York 1949; also *World Ethnographic Sample*, in: Moore, F. W. (ed.), *Readings in Cross-Cultural Methodology*, New Haven 1961.

<sup>39</sup> A survey in Germany showed that in 1977 there were more than two hundred machine readable data sets with quantified historical information; compare Bick, Wolfgang, et al., *Quantitative historische Forschung 1977*, Stuttgart 1977. (= *Historisch Sozialwissenschaftliche Forschungen*, Vol. 1). See also in the same series which is issued in cooperation with the International Association for Historical Social Research, *QUANTUM: Best, Heinrich, und Mann, Reinhard* (eds.), *Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung*, Stuttgart 1977, and Müller, *Die Analyse prozeß-produzierter Daten*.

<sup>40</sup> The development has gone furthest in the United States, and the best source to follow is the journal *Historical Methods Newsletter*, between 1968 and 1977 ten volumes. An example is Volume 9, Nos. 2 and 3 on one of the massive cases of quantitative history, the Philadelphia Social History Project. There is a very long tradition of a social science orientation with attention to quantitative data in France, the school of the *Annales*; cf. Iggers, Georg, *Die 'Annales' und ihre Kritiker*, in: *Historische Zeitschrift*, 219 (1974), pp. 579–608. The most direct impact on sociologists in English and German speaking countries can be attributed to the works



and the impact of social differentiation can only be studied by looking at other complex civilizations. It is indeed quite necessary to use historical complex civilizations for purposes of comparison in order to avoid a tendency in sociology to argue *post hoc propter hoc*. Bureaucracies, corporate associations, formalization of procedures are part of our daily life — but does that make them distinctive features of industrial societies? There is no other way to establish what is unique about industrial civilizations, and what is a feature of many complex societies, than to engage in historical comparisons. In this perspective the quantitative analysis of conditions during the Roman Empire at the time of the principat may contribute more to our understanding of contemporary industrial societies than yet another survey.

An example may help. In working on the sociology of vacations and tourism it is usual to assume that long-distance travel, weekend excursions, and the desertion of cities during the holiday season are phenomena unique to the very different industrial societies<sup>41</sup>. However, weekend traffic problems were part of life in the richer Greek cities, holiday desertion of cities was common amongst the bourgeoisie of classical Rome, and long distance travel institutionalized in several high civilizations such as Sumer, Persia, and Moghul India<sup>42</sup>. Several high civilizations even developed some infrastructure for travel, such as the road networks of ancient China or Persia or Rome, complete with a system of accommodations. However, at least one phenomenon appears to be unique to a modern civilization, namely the regular travel for pleasure only, while other travel such as the „Bildungsreise“ have been developed in other high civilizations.

Economic historians now inform us that production for markets is nothing unique to our industrial civilization<sup>43</sup>, nor is occupational specialization nor are election campaigns<sup>44</sup>. However, the differentiating out of economic activities appears to be a feature of our industrial societies, are a characteristic that to someone from a non-Western society gives our civilization a commercial flavor. In most cultures economic matters are subservient to political considerations, and political power is deemed a central goal and not economic well being. And in all other cultures econo-

of the brothers Tilly; Tilly, Charles, et al., *The Rebellious Century 1880–1930*, Cambridge/Mass. 1975; Shorter, Edward, and Tilly, Charles, *Strikes in France 1830–1968*, Cambridge/England 1974; Tilly, Charles, *The Vendée*, Cambridge/Mass. 1964; Tilly, Richard, *Popular Disorders in Nineteenth Century Germany*, in: *Journal of Social History*, 4 (1970). In Europe, Stein Rokkan in his many publications on nation-building has done more than any other individual scholar to further quantification of historical material for sociological analyses. An overview of the breadth of this development can be found in Flora, Peter, *Quantitative Historical Sociology*, in: *Current Sociology*, 23, No. 2 (1975).

<sup>41</sup> This is maintained e. g. in Scheuch, Erwin K., and Meyersohn, Rolf (eds.), *Soziologie der Freizeit*, Köln 1972, pp. 304–317. Some years later this opinion is revised in Scheuch and Scherhorn, *Soziologie der Freizeit und des Konsums*, pp. 115–147.

<sup>42</sup> Many details can be found in Casson, Lionel, *Reisen in der alten Welt*, London 1974.

<sup>43</sup> For Germany, urban history is the chief corrective for the previous inclination to mistake ideals for reality. See Kellenbenz, Hermann (ed.), *Zwei Jahrtausende Kölner Wirtschaft*, 2 vols., Köln 1975.

<sup>44</sup> Cf. Etienne, Robert, *La vie quotidienne à Pompéi*, Book no. 2, Chapter 2, Paris 1966.

mic relations between people who know each other are subservient to requirements and considerations of the social fabric<sup>45</sup>. Beyond economics, it may be possible that the generally distinguishing character of Western industrial societies is the sectorial rationality, the differentiating out of sector after sector from diffuse and multifunctional roles<sup>46</sup>.

However, such a statement may not last long in view of the many surprising findings of the history of our early industrial periods. Now we learn that not even early capitalism lived up to its reputation of mindless exploitation of helpless proletarians. Undoubtedly this occurred in the large industrial agglomerations, but in production and in living conditions on a smaller scale the employers cared not only for profit but also for their local reputation as human beings<sup>47</sup>. In addition to the comparison with other high civilization, the quantification of European data both of the late medieval period and of early industrialization are likely to be important contributions to our understanding of the distinguishing features of modern industrial societies.

This does not mean to just wait for the conclusions of historians, this requires data to be handled in ways that are usual in sociology, and for problem formulations that are sociological. This should be evident for what we believe to be characteristic for modern civilizations of the Western variety, namely the prevalence of sectorial rationality. It was already mentioned that in other civilizations our economic rationality is practiced in exchanges with outsiders, and this behavior is considered unfriendly. However, even with Western industrial societies there are limits to the extension of a specific economic rationality: We do not accept economic rationality between spouses, and between parents and children. For us, a really functioning family is based on communist sentiments, namely to each according to his needs and from each according to his abilities. In some areas such as sports, there is both sectorial rationality and diffuse standards, distinguishing the professional with a specific sectorial rationality from the amateur for whom sports has a diffuse meaning. Good research with the intention to specify sectorial rationalities requires the manipulation of historical material such as diaries or personal letters, looking for indications of value conflicts and for justifications of behavior. There is little hope that a historian would systematically look for indications of such aspects of behavior that are not part of the problem understanding of his discipline or of people themselves. In this sense there are many problems where sociologists cannot be consumers of conclusions from quantitative history but have to reexamine quantified historical material.

<sup>45</sup> This is a central theme in the research of Raymond Firth about the tribal cultures in the South Pacific, *Elements of Social Organization*, London 1951.

<sup>46</sup> This is the central notion in Scheuch, Erwin, K., *The Relationship of Government and Business to the Individual in Democratic and Totalitarian Systems*, in: *International Conference on the Unity of Science*, Vol. 4, International Cultural Foundation, New York 1978.

<sup>47</sup> See Stearns, Peter N., *Die Herausbildung einer sozialen Gesinnung im Frühindustrialismus — ein Vergleich der Auffassungen französischer, britischer und deutscher Unternehmer*, in: *Ludwig, Soziologie und Sozialgeschichte*, pp. 320–342.

## VI. The Importance of Descriptive Knowledge

However, quantitative history is of tremendous importance to sociologists in so far as it is an extension and more often a correction of social history. Students are still being tested by asking them to explain the *loi de contraction* by Emile Durkheim, and yet quantitative historical research shows that in all likelihood this *loi de contraction* is simply in error, is repeating what were the erroneous perceptions of eloquent contemporaries<sup>48</sup>. Provided we would hand on to our successors as the condition of public safety what our newspapers write, this would amount to a massive handing-on of misinformation; provided we were to hand on what magazines write about family life today, our successors would be better off without that information. However, many, many of the statements about daily life in the past are based on reports that are no more reliable than newspaper reports or the impression of contemporary intellectual gurus about our own industrial societies. Even if the guru or newspaper were correct about a condition or a change, they would be incompetent to characterize the diversity existing at this time and earning our societies the label „pluralistic systems“. Now that some preliterate cultures have been studied by more than one ethnologists we understand that even those relatively simple cultures have diversity, and that past ethnography reduced that diversity to an ideal type. It is reasonable that in historical societies there was no less diversity, that differences between actual behavior and official norms were common-place, and that an informal system paralleled official structure much as this is the case for our societies. Most social history is simply hopeless in these respects, and the only hope is the systematic analysis of large quantities of evidences of daily life in the past.

Was the exploitation of colonies a major cause for the economic development of France, or Germany? Was the French revolution caused by an intolerable pauperization of ordinary people? Was the middle of the 19th century in Germany a time when in economic controversies capitalists stood against labor? By now we know through quantitative history that the answer to all three questions is „no“<sup>49</sup> — and

<sup>48</sup> This is the conclusion of a number of quantitative studies in urban history, such as Hubbard, William H., *Der Wachstumsprozeß in den österreichischen Gross-Städten 1868–1910*, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, Special issue 16, pp. 386–418, and *Forschungen zu städtischer Haushaltsstruktur am Ende des 19. Jahrhunderts — Das GRAZ-HAUS-Projekt*, in: Conze, Werner (ed.), *Sozialgeschichte der Familie in der Neuzeit Europas*, Stuttgart 1977, pp. 283–291. Compare also Thernstrom, St., *The Other Bostonians*, Cambridge/Mass. 1976; Katz, M. B., *The People of Hamilton, Canada West*, Cambridge/Mass. 1975.

<sup>49</sup> For question no. 1 see Hochheimer, Albert, *Abschied von den Kolonien*, Zürich 1972; for question no. 2 compare Tilly, Charles, *Vendée*; for question no. 3 see Best, Heinrich, *Interessen-*

that is by no means unimportant for sociology. It will become even more evident to which degree we have based sociological statements on a social history that is becoming defunct.

For this author descriptive knowledge about the economy of the Roman Empire as it was furthered especially by historians in Oxford and in Princeton, became of great importance. My understanding of Rome was very much colored by the German historical tradition which concentrates on the turmoil period of the Roman Republic, and in the tradition of Theodor Mommsen understands this period as the corruption of republican ideals — which I now see as a perspective that is very much beside the point. By way of contrast British economic history has always emphasized the empire during its successful time — which after all is several hundred years. During this time the economic order was a variant, from a partial market economy to the centralized state socialism of Diocletian. During the whole time of the Roman Empire, the governments were unable to cure inflation and to establish a sound currency for any extended period — which, by the way, the Chinese Empire failed to do as well. Whether detailed regulations or market mechanisms: Nothing really worked.

And yet the Roman Empire failed to decay, while undoubtedly our systems would be mortally threatened if there would be inflation on the Roman scale over many decades. Being raised on Parsons I had believed that when interchanges are seriously upset there will be countervailing processes until the disturbances are corrected; the Roman Empire demonstrates that social systems can live with unsolved problems on a massive scale — provided there are redeeming features. In the case of the Roman Empire its performance as a political and legal order was obviously so impressive in comparison to other contemporary systems that the „Roman Way of Life“ was as successful an export article as the American Way of Life was after World War II.

Perhaps this is a general feature of highly differentiated societies: that they have „central problems“ but are at the same time able to live with them via redeeming features. More descriptive knowledge would help in translating this still very vague notion into a researchable question. However vague this notion, the descriptive material was already sufficient to correct the conventional wisdom in American sociology. Perhaps there will be reports based on quantitative history about other advanced civilizations that lasted hundreds of years without solving some central problem; perhaps these reports will inform us what the character of countervailing forces was.

It is especially quantitative history as the rewriting of conventional historiography, and as the extension of knowledge about forgotten eras that will have an important impact on sociology. One last example: We can expect important insights into the change of systems when the fifth century in Western Europe is being

analyzed. The notion of a Roman empire being overrun by screaming Barbarian hords bent on destruction is stark nonsense; Roman power did not collapse or was broken — it simply seeped away<sup>50</sup>. It is a story of desintegration and not of forceful destruction.

This is an exciting time both for historians and for social scientists. There is more than one way in which the discipline will benefit from the renewed encounter. Programmatic debates will have little utility in starting the development. Much the best way to aid this development is simply more empirical work.

<sup>50</sup> Sterzl, Anton, *Der Untergang Roms an Rhein und Mosel*, Köln 1978; also Ternes, Charles-Marie, *La vie quotidienne en Rhénanie Romaine (Ier—IVieme siecle)*, Paris 1972.

## The Emergence of Social Scientific History and the Analysis of Process-Produced Data: Some Introductory Remarks

If I am informed rightly, Arthur Schlesinger Jr. maintained at a similar discussion that all quantitative information is without great importance for the historian and that all important questions in history are of a qualitative nature. On the other hand, more than fifty years ago John Clapham stated in his simple and straightforward manner that economic history always ought to ask: how many, how large, how often — in other words that economic history always is quantitative in nature. We all know, of course, that just as political history cannot exclude *quanta* like the number of soldiers fighting in a war or the number of voters electing a political party to government, economic history also cannot be reduced to *quanta* since ideas, values, personalities and decisions play their part in economic life and cannot always be meaningfully subjected to measurements.

These are trivialities and I will not discuss them here. To approach the question which is put in the title of this panel — how the emergence of social scientific history is connected with the analysis of process-produced data — let me start with some personal recollections. More than 25 years ago, when I — without knowing what I was doing — started my first independent steps in social history I came across a collection of sources which nowadays could serve as a perfect starting point for a quantitative study of certain social strata and religious groups of eighteenth century southern Germany. As a twenty year old student of history I read many small autobiographies and „confessions“ of Suabian pietists which I found at the University library at Tübingen. The authors mainly were *petit bourgeois* and small independent peasants, living in the country-side or small towns in a similar life style, sharing a common belief, and worrying about the same problems. Most were fairly poor in material things but regarded themselves as rich spiritually, well prepared for an other-worldly life. They were eager to convince others of the rightness of their beliefs and attitudes, but did not care about politics, and even displayed a certain contempt for worldly distractions and power. Most of them had many children, and quite a few had a second wife, since the first often had died in childbirth. They were not mobile, neither geographically nor socially, but proud of staying at the place where God had put them and saw to it that their wives and children also would stay in their proper, *id est* inferior place. They worked hard and seemed to be good artisans and peasants but more important than *laborare* was *orare*. Praying and singing and praising God for all he had done and particularly for all the consolation he had given during a long day's hard work was what counted to

them. Occasionally there would be a merchant or a nobleman amongst them, equal as brother in faith, but distinct as a social being.

Fascinated, I wrote about two dozen life histories (essays of 8–20 pages each) to demonstrate what I thought was the typical, common feature of these men. Without really knowing, I wrote my first essays in social history and tried to sell them to a publisher hoping I could finance the rest of my studies through the sales of such a book. The lector of the publisher, an elegant lady of about 35 years, asked me into her office to talk about the manuscript. Although my writing was immature, she thought these stories were fascinating, but — above all — the book would not sell. Therefore, these first essays of mine are still unpublished, and if I have not lost them at one of my many moves since, they are still in the deepest corner of an old wardrobe where I keep my unfinished papers and materials. I earned my living, by the way, by writing about witchcraft trials and similar events for local newspapers. Unwittingly I was thus confronted with the difference between the *histoire des événements* and the *histoire de la mentalité* without knowing of the *Annales*. (Later I 'invented' modern qualitative and quantitative economic history for myself — a field which I never studied as a student). And when I wrote in my first published book, *Die Bildungswelt des deutschen Handwerkers um 1800. Studien zur Soziologie des Kleinbürgers im Zeitalter Goethes* (1955) that in the last instance every education is self-education („im Letzten ist jeder Autodidakt“), this early wisdom of my then 24 or 25 years was not only drawn from the 200 and odd autobiographies and letters I had read for this book but also from my own life experience.

If I had a student today who asked my advice after having found such sources himself — I wish I had such a student — I certainly would not recommend to him, if he wished to do serious scholarly work, to write a couple of impressionistic if impressive essays, but to read some sociology, study some similar dissertations, think about what could be done with such materials in the light of all this, learn some statistical procedures and perhaps a computer language, then start putting certain quantifiable features of his sources — like social origin, geographic mobility, age, number of children, social status, fortune etc. — on cards or tapes which finally could be processed either by hand or preferably by a computer. I would also recommend to try to quantify characteristics which can tell us about attitudes, e. g. certain key words and expressions, the contents of the prayers and songs, and the timetable of the daily life, in order to come up with a scientific-social history of this group of people in this particular time and region. Such a book would start with a theoretical framework, describe the sources and the methods used, be filled with lots of tables and appendixes, and culminate in some explanations and generalizations; it would probably be quite boring to everybody except those who want to write similar studies, while the book I mentioned before, the *Bildungswelt*, my more mature outcome of these first studies, is still fun to read 22 years after its publication.

Such is the emergence of social-scientific history and the analysis of process-produced data as I have witnessed it myself during the last quarter-century. What has changed and what is the situation now? The latter question will be discussed

by better trained and informed people during the next few days. I wish to concentrate in the second part of my remarks on the first one. What, if anything, has changed? Where and how has the availability of new techniques of research, the possibility to collect and process mass data of social historical relevance altered our knowledge? Does it affect our theoretical concepts, does it give deeper insights or does it only allow more precise answers for questions of the Clapham-type: how many, how large, how often? It seems obvious that the answer on this last part of the question is the most positive one. Data collection, and data processing have enabled scholars in many fields to know better, more exactly what we assumed anyway, but could not fully verify except perhaps by example. Whether we take historical demography or time series of foreign trade, exchange rates, prices and wages, business cycles, whether we think of the standard of living controversy or of voting behavior, all have benefited from the possibility to use mass data and to process them in manifold ways where formerly the working capacity of a single person or a group of researchers had set narrow limits. It may well be that in many cases we only know with certainty or somewhat more in detail what we knew before. Nevertheless, we must not underestimate the amount of scientific progress which can be promoted by such small steps.

But did we gain deeper insights? Certainly, the adoration of quanta has led to some overestimation of their cognitive value. There is a danger of making big efforts for *trivia*. Even otherwise very perceptive and subtle scholars may fall for the fascination of their new tools and report simple, self-evident results as important new insights, as Franklin Mendels has just criticized in a review of Emanuel Le Roy Ladurie's essays. Traditional historians often have this tendency in mind when they criticize the quantifiers. Not everything which can be quantified should be quantified. There is the very real danger that, as Albert Fishlow has put it, we know more and more about unimportant things but less about the important problems because they are too complex to be reduced to quantifiable terms. It would certainly be a mistake, therefore, to reduce social and economic history to quantitative history.

Probably negative is the answer, if we think of the results for the development of sociological theory or the basic concepts in social philosophy. I may be wrong, but I have the impression that most, if not all, the relevant sociological concepts have been developed before the age of the computer and that very little, if anything has been added through new research tools and techniques as they have become available during recent years. But in the long run, in the social sciences as well as in the physical science there will be a repercussion also in this area, because new methods and data are bound to promote new or, at least, more sophisticated concepts.

One fundamental change is already apparent. When I was a student I learned, among other now obsolete notions, that the difference between the social and natural sciences was that the social sciences cannot experiment, and do not have laboratories in which the researcher sets up controlled conditions. This is no longer true. Data processing allow us „to play around“ with data, to use different concepts



and techniques and to try the concept or data series which fits best. There is a lot of formal play which may just give intellectual satisfaction without opening up new insights. What does it mean for our understanding of say, business cycles and the reasons for their behavior, if we know that one type of a mathematically defined curve fits the data better than another one? Can we explain the oscillations of the economy any better? As the economist has to be careful not to fall for the money-veil that may obscure real movements in the economy, social historians have to be careful not to obstruct the sight of their subject through an elaborate formal apparatus. Here, as in problem-solving general, I believe in a pragmatic approach which allows selection of the proper method and technique by a kind of cost-benefit analysis of research in each case.

To conclude: It seems that the availability of techniques for data collecting and processing has widened the scope of scientific research in all kinds of socio-economic and socio-political history; it allows definite and testable small progress step by step and therefore narrows the gap between 'hard' and 'soft' sciences by making social history a bit harder than it used to be. But certainly these instruments only very slowly affect our basic concepts in the social sciences and cannot claim a monopoly at the expense of other, more traditional forms of interpreting and writing social and economic history.

## II. Analysis of Administrative and Census Data



## Problems and Opportunities in the Use of Individual and Aggregate Level Census Data

Ever since the first federal census was taken in 1790, scholars have used these data to try to understand American demographic and socio-economic development. Most of the nineteenth-century studies of census data were descriptive rather than analytical. During the past two decades, however, historians increasingly have used these nineteenth-century censuses to reconstruct American society from a more quantitative perspective.

Despite the increased usage of census data at both the aggregate and individual levels, very little effort has been made to assess the strengths and weaknesses of this source of information. Most scholars simply have used these data without really considering any of the problems inherent in them<sup>1</sup>. Other historians have altogether neglected this valuable source because they are unaware either of the information found in the censuses or of the analytical techniques available to investigate them<sup>2</sup>.

In this essay, I will consider very briefly some of the opportunities and problems in the use of American census data from the nineteenth century. I will first discuss some of the issues raised by the use of aggregate census data and then turn to the problems of using them at the individual level.

<sup>1</sup> This is particularly true of economic historians who tend to use census data without considering the possible biases in such data. For example, the recent studies of wealth inequality in nineteenth-century America have used census information without ascertaining exactly what is being measured by the questions relating to property. Soltow, Lee, *Men and Wealth in the United States, 1850-1870*, New Haven 1975; Soltow, Lee, *Patterns of Wealthholding in Wisconsin Since 1850*, Madison/Wisconsin 1971.

<sup>2</sup> Though most social historians now use census data in their analyses, many of them under-utilize the available information. Anthony Wallace's recent study of an industrial community uses census information as part of its analysis, but it fails to use those data to explore in more depth the lives of the inhabitants of that community. Wallace, Anthony F. C., *Rockdale: The Growth of an American Village in the Early Industrial Revolution*, New York 1978.

## I. Aggregate Use of Census Data

Scholars from very different perspectives have utilized aggregate census data to investigate our past. Political historians have studied patterns of electoral behavior at both the state and country levels using the federal censuses in conjunction with the national and state election returns. Fertility differentials and trends have been studied by demographic historians while economic historians have analyzed nineteenth-century economic development. Finally, social historians are now beginning to investigate such issues as school attendance, literacy, and urban development.

The first American censuses were gathered primarily for political purposes. As part of the „Great Compromise“ of the Convention of 1787, each state received equal representation in the Senate while the House of Representatives was apportioned on the basis of the population of each state. Therefore, it was necessary to provide for some mechanism for counting the population. As a result, a census of the population was instituted for 1790 and held at ten-year intervals thereafter<sup>3</sup>.

The first federal census was not very comprehensive. Information was gathered only on six items:

- 1) The name of the head of the household
- 2) The number of free white males of 16 years and upwards, including the head of the household
- 3) The number of free white males under 16 years
- 4) The number of free white females, including the head of the household
- 5) The number of all other free persons
- 6) The number of slaves

It was not until 1830 that regular census schedules were printed and distributed to the census marshalls. Up to that time, each census enumerator simply made their own forms. Combined with the fact that the position of a census marshall was often reserved for rewarding politicians, it is not surprising that the early censuses were not as carefully and accurately carried out as their modern counterparts.

Information for the federal censuses from 1790 to 1840 were collected only at the household level. Therefore, it is impossible to obtain individual level data for any of these years. Beginning with the federal census of 1850, however, the individual became the object of enumeration — a major innovation and improvement in the quality of the data. Furthermore, by that date the questions asked of the population had been greatly expanded to include additional information on such issues as occupation and wealth. Altogether, six separate census schedules were used in 1850 to obtain data on population (both free and slave), agriculture, industry,

<sup>3</sup> For a useful introduction to the development of the federal censuses, see Wright, Carroll D., and Hunt, William C., *The History and Growth of the United States Census*, Washington D.C. 1900.

mortality, and social statistics (i. e. schools and colleges, libraries, newspapers and magazines, religion, crime, poverty, and wages). Thus, during the course of the nineteenth century, the federal censuses became more detailed and comprehensive both in the type of information they solicited and the manner in which it was recorded.

Much of the census information gathered during the nineteenth century was aggregated at either the state or county level and the results were published by the federal government. These published aggregate data have become the basis of the ecological analyses of census materials. Analyses of these data are greatly facilitated by their availability at the state and county levels in machine readable form through the Inter-University Consortium for Political and Social Research<sup>4</sup>.

There are problems, however, with the manner in which the census data were aggregated. The unit of aggregation is usually either the state or the county. But state and county level data often mask significant socio-economic variations within units. Some scholars have used township data for their analyses even though it has meant going back to the original census manuscripts and reaggregating the data or going to some other source such as the state censuses which sometimes are aggregated at the township level. Simply to assume that the proper level of analysis is the state or county because of the lack of data at the township level is a mistake. Some of the most interesting analyses in fields such as educational development need to be examined at the township rather than county level since local variations in the pattern of school attendance or school expenditures often can be quite sizable within counties.

Another problem of using the available aggregate census data is that the units are of such varying size. For example, the population of the largest state in 1850, New York, was 3,048,325 people, while that of the smallest state, Florida, was only 47,203 people. Similarly, whereas the county of New York (New York) had a population of 515,547 people, there were only 79 inhabitants in the county of Clarke (Iowa) in 1850. Anyone doing ecological analysis has to decide whether to use these units as equivalent to each other or to weight them by some factor such as their population size. The correct procedure, of course, depends on the model being tested by our analysis. In most historical work, analysts have treated all units as equivalent and therefore have not weighted them. Generally, this is a conceptually defensible procedure, but sometimes we do encounter serious problems if one of the smaller units has an extreme value.

Let me illustrate this latter point by referring to a study of white fertility ratios at the state level in ante-bellum America. Colin Forster and G. S. L. Tucker analyzed fertility differentials using all the states and territories in the United States between 1800 and 1860. The total number of units in their analysis was always quite small since there are only a limited number of states and territories in this period. Thus, in 1860 there were only thirty-four states and seven territories in the

<sup>4</sup> For a catalogue of the available machine-readable data at the Inter-University Consortium, see Inter-University Consortium for Political and Social Research, *Guide to Resources and Services*, 1977-1978, Ann Arbor/Michigan 1978. These Guides are updated annually.

United States (not counting the District of Columbia)<sup>5</sup>. As a result, one has to be particularly careful that none of the states or territories have extreme values in any of their variables since this might distort the entire analysis.

In their analysis of 1860, Forster and Tucker used all of the states and territories. Unfortunately, this meant including the Dakota territory which has a white population of only 2,576 in that year. Since the Dakota Territory also had a very low white refined fertility ratio (1,157) and an unusually high white sex ratio (1,710), it seriously skewed their overall results. Rather than have the entire analysis distorted by such an extreme case, they should have either eliminated all of the territories or weighted their states and territories by population size so that their results would be less affected by a new and unsettled area that had such unusual characteristics compared to the other states<sup>6</sup>.

The above example also highlights one of the major problems of using aggregate census data at the state level — the small number of cases available. With only thirty-four states in 1860, it is very difficult to use any elaborate regression equation since the degree of freedom lost by each additional independent variable restricts our analysis. Furthermore, any extreme value in any of the cases, whether they be for small or large units, can greatly affect the results since there are so few cases in the analysis. Finally, it is almost impossible to analyze variations within regions of the country using state level data because of the small number of states within any region. Unfortunately, there are examples in the historical literature of correlation analysis being done with such small number of states<sup>7</sup>.

In analyzing aggregate census data over time, a further complication arises due to changes in the boundaries of the units. This is particularly true at the county level in the nineteenth century. In many of the newer states, when the population of the counties increased, they were subdivided into smaller units. As a result, direct county to county comparisons over time are difficult for nineteenth-century America.

The use of aggregate census data is also hindered because census units do not always coincide with political units. For example, congressional districts in nineteenth-century America sometimes split counties as the state legislature was more concerned about the characteristics of the voters within each district than about keeping county lines intact. As a result, it is difficult to create files of political and census data at the congressional level.

<sup>5</sup> Forster, C., and Tucker, G. S. L., *Economic Opportunity and White American Fertility Ratios, 1800–1860*, New Haven 1972.

<sup>6</sup> For a critique of their analysis, see Vinovskis, Maris A., *Socio-Economic Determinants of Interstate Fertility Differentials in the United States in 1850 and 1860*, in: *Journal of Interdisciplinary History*, 6 (1976), pp. 375–396; Vinovskis, Maris A., *Recent Trends in American Historical Demography: Some Methodological and Conceptual Considerations*, in: *Annual Review of Sociology*, 4 (1978), pp. 603–627.

<sup>7</sup> Some of the analyses at the state level in Yasukichi Yasuba's work are done with only ten or fifteen cases. Yasuba, Yasukichi, *Birth Rates of the White Population in the United States, 1800–1860: An Economic Study*, Baltimore 1962.

In addition to considering the problems of the comprehensiveness of the census questionnaires as well as the size, number, and comparability of the aggregate census units, we also need to evaluate the quality of the census returns. Most scholars using historical censuses have not paid sufficient attention to the accuracy of the censuses. Usually they simply assume that these data are accurate or that if they are under-enumerated, the degree of under-enumeration is relatively uniform so that comparisons across units are still meaningful.

The quality of census data varies not only by the particular census which is being used, but also by what issue is being considered. Generally, the later censuses are more accurate than the earlier ones since more standardized procedures were introduced for gathering and processing the data. There are some censuses, however, such as that of 1870, which are considered to be inferior in quality to their predecessors.

The data on the general characteristics of the members of a household are considered to be reasonably accurate, though it is clear that some households have been omitted entirely — particularly those which were the most transient. Though scholars are still concerned about the accuracy of these data, most of them are quite willing to accept and use them in their investigations<sup>8</sup>. Other areas of census inquiry are so clearly unreliable that everyone agrees that these data should not be used in further analyses. For example, the data on the insane population are clearly under-enumerated to such a large extent that even contemporary observers such as Edward Jarvis argued that they were worthless. The nineteenth-century federal census marshalls were negligent in collecting the information on insanity and the families they interviewed were often reluctant to admit that a member of their household was insane<sup>9</sup>.

Though there is a consensus among scholars on the relative reliability of some categories of census information, there is considerable difference of opinion on other areas such as mortality information. Most demographers and economists continue to use mortality data from the census of 1850 even though other scholars have argued that these data are so badly under-enumerated that any conclusions based on them are highly suspect. Some defenders of the use of these mortality data argue

<sup>8</sup> A few historians have tried to consider the accuracy of the census data. For example, several scholars have investigated the accuracy of age reporting in the federal censuses. Knights, Peter R., *Accuracy of Age Reporting in the Manuscript Federal Censuses of 1850 and 1860*, in: *Historical Methods Newsletter*, 4, No. 3 (June 1971), pp. 79–83; Hammarberg, Melvyn, *The Indiana Voter: The Historical Dynamics of Party Allegiance During the 1870's*, Chicago 1977, pp. 210–217.

<sup>9</sup> For a discussion of the under-registration of the insane in the early federal censuses, see Grob, Gerald N., *Edward Jarvis and the Federal Census*, in: *Bulletin of the History of Medicine*, 50 (Spring 1976) pp. 4–27; Rosenkrantz, Barbara G., and Vinovskis, Maris A., *The Invisible Lunatics: Old Age and Insanity in Mid-Nineteenth-Century Massachusetts*, in: Spicker, Stuart F., et al. (eds.), *Aging and the Elderly: Humanistic Perspectives in Gerontology*, Atlantic Highlands/N.J. 1978, pp. 95–125.



that even if these data are under-enumerated, the relative under-enumeration is about the same so that these data are still appropriate for comparative purposes. But there is no reason to believe that mortality data are under-enumerated at the same rate throughout the country or among different segments of the population. In fact, the superintendent of the census, as well as other nineteenth-century scholars, argued that the mortality returns were so badly and unevenly under-enumerated that they were of little value even for comparative purposes.

The federal census of 1850 furnishes the first instance of an attempt to obtain the mortality during one year in all of the States of the Union, and had there been as much care observed in the execution of the law as was taken in framing it, and in the preparation of the necessary blanks, a mass of information must have resulted relating to the sanitary condition of the country, attained as yet in no other part of the world . . . The varying ratios between the States, as drawn from the returns, show not so much in favor of or against the health of either, as they do, in all probability, a more or less perfect report of the marshalls. Thus it is impossible to believe Mississippi a healthier State than Rhode Island, etc.<sup>10</sup>

Thus, one of the major criticisms that can and should be raised against many of the scholars using nineteenth-century aggregate censuses is that not enough attention has been paid to the quality of the data. Much of the recent work in this area is quite sophisticated from a statistical perspective, but very elementary in terms of considering the possible biases in the data. One simply cannot use nineteenth-century census materials as if they were as accurate as those produced by the U. S. Bureau of the Census today.

One other point needs to be mentioned in regard to the accuracy and representativeness of census data. Even if the census data are accurate, it does not mean that they are typical or representative of that decade. Almost none of the studies using census data have adequately tested the possibility that the year in which the census was taken was atypical in terms of the variable being investigated. This is particularly hazardous for historical data since there were often much larger and more frequent fluctuations in demographic and socio-economic variables in the past than today.

Again, let me illustrate this point by using mortality data from the federal census. One of the major life tables for nineteenth-century America is the Jacobson Life Table. It is based on census mortality data for Massachusetts and Maryland in 1850. The data were originally assembled and converted to life expectancies by L. W. Meech and recalculated by Paul Jacobson a hundred years later<sup>11</sup>. The Jacobson Life Table has been accepted by virtually all scholars as the definitive estimate of life expectancy in mid-nineteenth-century America.

<sup>10</sup> U. S. House of Representatives, *Mortality Statistics of the Seventh Census of the United States*, House Executive Documents No. 38, 33rd Congress, 2nd Session, Washington D. C. 1855, p. 8.

<sup>11</sup> Jacobsen, Paul H., *An Estimate of the Expectation of Life in the United States in 1850*, in: *Milbank Memorial Fund Quarterly*, 35 (1957), pp. 197-201.

The question we need to ask is whether the period June 1, 1849–June 1, 1850 was a typical year in terms of mortality. Naturally, since the census itself covers only one year, it is not of much help in answering this question. Fortunately, Massachusetts had established a state vital registration system in 1843 so that we have annual mortality information.

It turns out that 1849 was a very unusual year in terms of mortality because of the outbreak of cholera which greatly inflated death rates. Thus, whereas the expectation of life for Massachusetts males at birth was 39.4 years in 1849, in 1850 it was 46.0 years, and in 1851 it was 43.0 years. Clearly, the Jacobson Life Table exaggerated the extent of mortality because it is based on census data from an unusually unhealthy year<sup>12</sup>.

However, earlier we argued that the mortality data from the censuses were probably under-enumerated; perhaps the exaggeration of mortality due to the cholera epidemic was compensated by the under-enumeration of the mortality data. Though this is plausible, it does not appear to be true. When we analyze the Massachusetts mortality data in more detail, it still appears that the Jacobson Life Table for 1850 greatly exaggerates the extent of mortality in that state for most years during the antebellum period<sup>13</sup>. Thus, we need to be extremely cautious in accepting census estimates of phenomena without first checking to see whether that year might have been atypical.

Since most studies of aggregate census data rely on the published summaries of the censuses, we also need to consider the accuracy of these published summaries. If the published summaries of the census are not an accurate tally of the individual census returns, then studies at the aggregate level will be using incorrect data. Though relatively little attention has been paid to this issue, two recent studies of the accuracy of the printed census summaries suggest that there is a serious problem for the use of at least some aggregate data.

Edward Muller has investigated the reporting of the populations of smaller towns in the federal censuses through 1870<sup>14</sup>. Although the populations of most towns larger than 2500 were published regularly after the 1810 census, he found that the population of smaller towns were either under-reported or the towns were entirely omitted from the published census summaries. Muller found that the pro-

<sup>12</sup> For a critique of the Jacobson Life Table, see Vinovskis, Maris A., *The Jacobson Life Table of 1850: A Critical Re-Examination from a Massachusetts Perspective*, in: *Journal of Interdisciplinary History*, 8, No. 4 (Spring 1978), pp. 703–724.

<sup>13</sup> For additional analyses of nineteenth-century mortality patterns, see Jaffe, A.J., and Lourie, W. I., *An Abridged Life Table for the White Population of the United States in 1830*, in: *Human Biology*, 14 (1942), pp. 352–371; Yasuba, Birth Rates, pp. 86–96; Thompson, Warren S., and Whelpton, P. K., *Population Trends in the United States*, New York 1933, pp. 228–240; Vinovskis, Maris A., *Mortality Rates and Trends in Massachusetts Before 1860*, in: *Journal of Economic History*, 32 (1972), pp. 184–213.

<sup>14</sup> Muller, Edward K., *Town Populations in the Early United States Censuses*, in: *Historical Methods Newsletter*, 3, No. 2 (March 1970), pp. 2–8.

portion of populations of all towns (over 100 inhabitants) in forty-one counties in southwestern Ohio and southeastern Indiana from 1820 to 1860 ranged in the published censuses from nineteen percent to seventy-three percent (the average for the period was forty-one percent). In other words, studies of small town populations based on the published nineteenth-century censuses would be very inaccurate.

The accuracy of the published mid-nineteenth-century manufacturing returns for Wisconsin have been studied by Margaret Walsh<sup>15</sup>. She found that in only 19.8 percent of the cases did the figures compiled from the manuscript manufacturing census coincide with those in the printed census summaries. In most instances the differences were only slight. However, in about one third of the cases at the county level, the differences between the manuscript and printed censuses were at least of the magnitude of ten percent. Fortunately, at least for the case of Wisconsin in 1850 and 1860, the errors at the state level were quite small as many of the errors at the county level cancelled each other out. Thus, Walsh concluded that the major advantage of using the manuscript census of manufacturing rather than the printed summary is that the former provides more accurate information at the county level.

Both of these examples illustrate the problems of using the aggregate census returns without checking the original manuscript returns. Though for most variables we would not anticipate a very large difference between the manuscript and the printed returns, one should at least consider the possibility. Unfortunately, the logistical problems entailed in such a verification procedure are enormous — especially for studies which focus on the country as a whole rather than just one state. Nevertheless, we should at least be aware of this potential source of error and how it might affect our analyses — especially at the county level.

After we have dealt with the issues of the level of the analysis and the quality of the data, we can turn to the analysis of the materials. In many respects, the analysis of aggregate census data is the same as that of dealing with any other data sets. There are, however, some important and interesting conceptual and statistical issues raised by the way in which historians have dealt with aggregate censuses that should be discussed.

One of the most difficult problems in any analysis is to develop from the available data the appropriate indices to test our hypotheses. This is particularly difficult with nineteenth-century aggregate census data since we do not have good information for many of the socio-economic variables we would like to include in the analysis. Historians have sometimes developed indices that are inappropriate or inadequate measures of the concepts being investigated.

For example, economic historians such as Yasukichi Yasuba and others have argued that the increasing scarcity of readily available farmland in nineteenth-century America accounts for the decline in fertility among the rural white popula-

<sup>15</sup> Walsh, Margaret, *The Census as an Accurate Source of Information: The Value of Mid-Nineteenth Century Manufacturing Returns*, in: *Historical Methods Newsletter*, 3, No. 4 (September 1970) pp. 2–13.

tion during these years<sup>16</sup>. As his measure of the availability of farmland, Yasuba calculated the number of persons per 1000 arable acres. But his index was based on the cropland in 1949 and properly has been criticized for reflecting the levels of twentieth-century farming technology and practices rather than nineteenth-century agricultural potential.

The most recent effort by Forster and Tucker calculates the number of white adults per farm, using the white adult farm population in the census year under investigation and the number of farms in 1850, 1860, and 1880. Their index has the advantage of reflecting nineteenth-century farming conditions and practices more accurately than Yasuba's measure<sup>17</sup>.

However, even Forster and Tucker's index of land availability leaves much to be desired. At the state level, an index of white adults per farm is highly correlated with the percentage of the population engaged in nonagricultural occupations and with the percentage of the population in urban areas. Therefore, we cannot be sure whether the high correlation between the white adult-farm ratio and the white refined fertility ratio is due to the availability of farms, to the percentage of the population in nonagricultural occupations, or to the percentage of the population living in urban areas.

The number of white adults per farm is not only an ambiguous measure of land availability in terms of being highly correlated with other indices, but it is also conceptually weak in that it does not reflect the relative cost of establishing a farm household. When economists speak of the availability of farms, they are in effect considering the relative costs of establishing a farm. Forster and Tucker's measure of agricultural opportunity implicitly treats all farms as equally priced, though in reality there are wide differences in the costs of farms in ante-bellum America. Thus, to take an extreme example, the average value of a farm in 1860 in Kansas was \$1,179, whereas the average value of a farm in Louisiana was \$11,818 in the same year. Surely it was more difficult for a young man to purchase a farm in a state such as Louisiana than in Kansas<sup>18</sup>.

Historians have used a variety of statistical methods to analyze aggregate census data. One of the most common procedures among political historians is the use of relatively homogenous units as the basis of their analysis. Thus, communities with a high percentage of German Lutherans are used as an indicator of how German Lutherans were voting in a given election<sup>19</sup>.

Though this method of analysis has been used extensively in American political history, it has been properly criticized for using very atypical groups of immigrants

<sup>16</sup> Yasuba, *Birth Rates*.

<sup>17</sup> Forster and Tucker, *Economic Opportunity*.

<sup>18</sup> Vinovskis, *Determinants*; Vinovskis, Maris A., *Demographic History and the World Population Crises* (Chester Bland-Dwight E. Lee Lectures in History), Worcester/Mass. 1976.

<sup>19</sup> Jensen, Richard J., *The Winning of the Midwest: Social and Political Conflict, 1888-1896*, Chicago 1971; Kleppner, Paul, *The Cross of Culture: A Social Analysis of Midwestern Politics, 1850-1900*, New York 1970.

— those who chose to or were forced to live together. There is no reason to assume that German Lutherans living in a community that was composed almost entirely of their countrymen voted the same way as those German Lutherans living in more heterogeneous communities. Furthermore, there is an implicit assumption in these studies that ethnicity or religion are the major determinants of voting behavior rather than some other variable such as occupation or wealth. By using relatively homogeneous communities as indicators of voting behavior rather than applying multivariate techniques of analysis to all of the communities in that area, these studies have under-utilized the type and variety of factors that should be considered in any mass voting analysis<sup>20</sup>.

Another common procedure for analyzing aggregate census data is to cross-tabulate the data. Unfortunately, this method usually permits us to study only two variables at a time. If we try to use cross-tabulation techniques to control for a third or fourth variable by further subdividing our data, the number of entries in our cells often becomes so small as to hinder our analysis. As a result, historians are increasingly turning to multivariate techniques such as regression analysis<sup>21</sup>.

In most instances, multiple regression analysis is preferable to cross-tabulation in the analysis of aggregate census data. But one must be very careful in using regression analysis or any other statistical procedures not to violate the fundamental assumptions on which they are based. Historians have not always paid sufficient attention to these problems. For example, multiple regression analysis assumes that the independent variables are independent of each other. Naturally, in any real life situation, the independent variables will be interrelated, but usually not at such a high level as to invalidate the analysis. There are situations where the independent variables are so highly correlated, however, that we encounter the problem of multicollinearity. There are examples in the historical literature of where scholars have included independent variables in their multiple regression analyses that were too highly correlated with each other<sup>22</sup>.

Finally, we should briefly mention one of the major debates among scholars using aggregate census data — the issue of inferring individual characteristics from ecological data. Most historical studies using aggregate census have made inferences

<sup>20</sup> For a good critique of these studies, see Wright, James E., *The Ethnocultural Model of Voting: A Behavioral and Historical Critique*, in: *American Behavioral Scientist*, 16, No. 5 (May/June 1973), pp. 653–674.

<sup>21</sup> Some historians have tried to defend the use of cross-tabulation of data instead of employing multivariate techniques. For an interesting though somewhat misleading exchange on this issue, see Katz, Michael B., *Who Went to School?*, in: *History of Education Quarterly*, 12 (Fall 1972), pp. 432–454; Denton, Frank, and George, Peter, *Socio-Economic Influences on School Attendance: A Study of a Canadian County in 1871*, in: *History of Education Quarterly*, 14 (Summer 1974), pp. 223–232; Katz, Michael B., *Reply*, op. cit., pp. 233–234; Denton, Frank, and George, Peter, *Socio-Economic Influences on School Attendance: A Response to Professor Katz*, op. cit., (Fall 1974), pp. 367–369; Calhoun, Daniel H., *Letter to the Editor*, op. cit., (Winter 1974), pp. 545–546.

<sup>22</sup> E. g. Forster and Tucker, *Economic Opportunity*.

only at the ecological level of the township, county, or state. But during the past five years, some historians, particularly those in the area of mass politics, have tried to infer how individuals voted on the basis of aggregate census and electoral data.

The origin of this debate goes back to W. S. Robinson's influential article on the ecological fallacy in which he warned that ecological correlations are not the same as individual level correlations<sup>23</sup>. Since then, several scholars such as Leo Goodman and others have tried to develop methods of using ecological regressions to estimate individual level behavior<sup>24</sup>. In the area of historical analysis, Allan Lichtman and Morgan Kousser have applied these techniques to the study of electoral behavior<sup>25</sup>.

It is possible to estimate individual characteristics from ecological data under certain very limiting assumptions. Thus, if we are willing to assume that the relationship between our variables will be constant across all of our units, it is possible to make reasonable inferences about the behavior of individuals from aggregate census returns. However, it is rarely the case that the relationship between any two variables will be constant across all units — especially since there is a tendency for individuals either to move to areas with certain characteristics or to adjust their behavior in their new environment. In other words, though it is possible in certain situations to make inferences about individual level behavior from aggregate census returns, it usually presupposes the type and extent of knowledge about these variables that we simply do not have. As a result, though the use of ecological regression analysis to make estimates of individual characteristics has stimulated better efforts to specify variables and their pattern of interaction, it is unlikely that we can safely use such techniques to make firm estimates for much of the historical data available from the aggregate censuses<sup>26</sup>.

<sup>23</sup> Robinson, William S., *Ecological Correlation and the Behavior of Individuals*, in: *American Sociological Review*, 15 (1950), pp. 351–357.

<sup>24</sup> Goodman, Leo A., *Ecological Regressions and Behavior of Individuals*, in: *American Sociological Review*, 18, No. 6 (December 1953), pp. 663–666; Goodman, Leo A., *Some Alternatives to Ecological Correlation*, in: *American Journal of Sociology*, 64, No. 5 (March 1959), pp. 610–625.

<sup>25</sup> Lichtman, Allan J., *Correlation, Regression and the Ecological Fallacy: A Critique*, in: *Journal of Interdisciplinary History*, 4 (1974), pp. 417–433; Lichtman, Allan J., *Critical Election Theory and the Reality of American Presidential Politics, 1916–1940*, in: *American Historical Review*, 81, No. 2 (April 1976) pp. 317–349; Kousser, J. Morgan, *Ecological Regression and the Analysis of Past Politics*, in: *Journal of Interdisciplinary History*, 4, No. 2 (Autumn 1973), pp. 237–262; Kousser, J. Morgan, *The Shaping of Southern Politics: Suffrage Restriction and the Establishment of the One-Party South, 1880–1910*, New Haven 1974.

<sup>26</sup> For further discussions of the problem of inferring individual behavior from ecological data, see Dogan, Matthei, and Rokkan, Stein (eds.), *Quantitative Ecological Analysis in the Social Sciences*, Cambridge/Mass. 1969; Orcutt, Guy H., et al., *Data Aggregation and Information Loss*, in: *American Economic Review*, 58, No. 4 (September 1968), pp. 773–787; Feige, Edgar L., and Watts, Harold, *An Investigation of the Consequences of Partial Aggregation of Micro-Economic Data*, in: *Econometrica*, 40, No. 2 (March 1972), pp. 343–360; Hammond, John L., *Two Sources of Error in Ecological Correlations*, in: *American Sociological Review*, 38, No. 6 (December 1973), pp. 764–777; Wasserman, Ira M., and Segal, David R., *Aggregation Effects*

## II. Individual Use of Census Data

Much of the recent interest in the use of the censuses has been focused on the individual level data available on the manuscript censuses. As we noted earlier, it was not until 1850 that the census began to enumerate their data on the basis of the individual rather than the household. Furthermore, the manuscript census of 1900 has not been available to scholars until very recently and then so only under conditions of restricted access in terms of where one can use them (either in Washington D. C. or in one of the regional depositories). Yet the use of individual level census information has become a major activity of American social historians today and is likely to grow in importance in the near future as historians begin to utilize these individual level returns even more effectively.

Most of the aggregate census studies have been done at the national or regional levels. Only a few have focused on just one state and even fewer have studied only a portion of a state. Studies using individual level census data, on the other hand, have almost always been done on a small geographic area — usually a town or city. In fact, most of the individual level census analyses have been done by urban historians who have studied a particular community such as Boston or New York City<sup>27</sup>.

The almost exclusive focus of individual level analyses on single urban communities has been rather unfortunate in at least two respects. First, very little effort has been made to design research projects to include different types of urban and industrial development for comparative purposes. Second, the reliance on only urban areas has made it impossible to separate analytically the effects of urban development from more general changes within that society. In other words, by not having any rural control areas in their analyses, for example, researchers cannot be certain whether the changes experienced by any group over time within a city are the result of the impact of urbanization on their lives or the consequence of more general developments within that society as a whole.

There are alternatives to the single community focus. One such example is the study of eight Essex County (Massachusetts) communities in 1860 and 1880. This study, initiated by Tamara Hareven and myself, was designed to study the interaction between community structure and family life in one small area. By restricting the analysis to eight communities within one county, we minimized any regional differences in our analysis. We do not, of course, claim that this particular county is in any way typical or representative of the country as a whole, but only that the di-

in the Ecological Study of Presidential Voting, in: *American Journal of Political Science*, 17, No. 1 (February 1973), pp. 177–181; Shively, W. Phillips, 'Ecological' Inference: The Use of Aggregate Data to Study Individuals, in: *American Political Science Review*, 63, No. 4 (December 1969), pp. 1183–1196.

<sup>27</sup> A notable exception to this approach is Blumin, Stuart M., *The Urban Threshold: Growth and Change in a Nineteenth-Century American Community*, Chicago 1976.

versity of activities within it permit interesting and useful analyses to be done at both the community and household levels<sup>28</sup>.

The communities selected within Essex County provided a variety of experiences and opportunities for their inhabitants. We chose three large, urban areas with different types of economic activity — Lawrence, Lynn, and Salem. Lawrence was a new city developed around the textile industry while Lynn was an old city dominated by the shoe industry. Salem was an old commercial center that only became heavily industrialized after the Civil War. In addition, we selected five rural areas — Boxford, Hamilton, Lynnfield, Topsfield, and Wenham. Though all of these communities were small in term of their population size, they also varied in their economic activities. Some, like Boxford, were almost totally agricultural, while others, like Lynnfield, had a sizable proportion of their population already engaged in the shoe industry.

We must approach the use of individual level census data within the perspective of their broader environment. The lives of women living in nineteenth-century Pittsburgh, for example, were quite different than those of women in Lawrence because of the different employment opportunities available to them in those two communities. By consciously trying to select areas of varying population size and economic activity, we can develop a more useful setting for exploring individual level census information. Studies of individuals in the past which do not even consider the impact of the nature of the community in which these people lived are badly flawed. Urban historians in particular would greatly enhance their analyses by trying to develop more controls within their research projects in order to permit them to examine the relationship between family life and community setting.

One might argue that we should forego either the single urban focus or even the broader approach suggested by the Essex County project for a national sample of households from the manuscript censuses in 1850, 1860, 1870, 1880, and 1900 (the manuscript returns for 1890 were destroyed in a fire). There is considerable merit in the idea of a national sample from the nineteenth-century censuses — particularly if the sample sizes were large enough to reflect rural-urban as well as regional differences. This approach, however, does not negate the need for identifying individuals and households within the context of the communities in which they lived since we could attach some of that information even in the national sample.

Though a national sample from the nineteenth-century manuscript censuses is a good idea and should be given very high priority in the near future, it will not eliminate the need for more indepth analyses at the local level. Local studies can deal with the interaction of family life and community opportunities in a way is difficult to study at any other level. On the other hand, future local community studies need to be designed so that they are more than just another example of a Newburyport or Boston.

<sup>28</sup> For studies using the Essex County data, see Hareven, Tamara K., and Vinovskis, Maris A. (eds.), *Demographic Processes and Family Organization in Nineteenth-Century American Society*, Princeton 1978; Hareven, Tamara K. (ed.), *Family Transitions and the Life Course*, New York 1978.



So far we have discussed the type of setting from which the individual level data should be gathered — ranging from a single urban community to the nation as a whole. Now we will turn to the issue of how these data should be assembled once we have selected the appropriate geographic areas for analysis.

In most studies, the unit of analysis is the household. The data on the household are assembled either by having a separate card for each member of the household or by simply summarizing some of the household characteristics from the individual returns. Though the latter method is usually much less expensive in terms of coding time and keypunching expenses, it is also more apt to produce errors and makes it very difficult to use those data for some other purpose. Thus, one of the pioneering studies from the manuscript census is Merle Curti's study of Trempealeau County (Wisconsin) which assembled its data by summarizing the household information from the census<sup>29</sup>. Unfortunately, Curti and his associates were not particularly interested in many of the family issues that historians today find so intriguing. As a result, it is virtually impossible, short of going back to the original manuscript census returns, to reanalyze the Trempealeau County data for such issues as the pattern of school attendance or marital fertility. If these data had been organized along individual as well as household lines, they should be of much greater value to us today.

Some of the studies based on individual census data collect information on all of the inhabitants in the area. Many of the others sample the data. Unfortunately, many of the historical studies using some form of sampling are very badly flawed either in the way the data were sampled or in the size of the sample itself. For example, some studies have drawn samples from very different sources but have treated them as comparable anyway. Stephen Thernstrom's analysis of social mobility is based on samples from five different sources — the 1880 manuscript federal census returns, Boston's 1910 marriage license registers, the 1930 Boston birth records, the Boston City Directory for 1958, and Edward Laumann's survey sample of the suburban communities of Cambridge and Brighton<sup>30</sup>. These are very different sources of data and are not comparable to each other. Nevertheless, Thernstrom's analysis utilizes all five rather interchangeably without adequately considering the possible biases that may have been introduced.

The second major sampling problem characteristic of many of these studies is the small number of households sampled. Historians have been almost totally unaware and unconcerned about the problem of errors introduced by using samples rather than total populations. For instance, a recent study of southern Michigan drew random samples for Detroit of 70 households in 1850 and 102 households in 1880<sup>31</sup>. The analysts then proceeded to investigate the data for Detroit by the

<sup>29</sup> Curti, Merle, *The Making of an American Community: A Case Study of Democracy in a Frontier County*, Stanford 1959.

<sup>30</sup> Thernstrom, Stephen, *The Other Bostonians: Poverty and Progress in the American Metropolis, 1880–1970*, Cambridge/Mass. 1973.

<sup>31</sup> Bloomberg, Susan E., et al., *A Census into Nineteenth-Century Family History: Southern*

occupation of the head of the household for the native-born and the foreign-born populations. The sampling error for many of their calculations is so high as to make many of their conclusions statistically unreliable. Unfortunately, this misuse of census data by drawing samples that are much too small is a very common problem among American historians today.

Even if the sample sizes selected are adequate for the purposes of the study, one needs to decide which sampling procedure to follow. Though many of these historical studies claim to be based on a random sample of households from the manuscript census, most of them are really based on a systematic sample. That is, most of the studies have sampled every  $n^{\text{th}}$  household in the manuscript census rather than giving each household an equal chance of being selected at random. The use of a systematic sample is a defensible procedure for most purposes, but the manner in which it has been done by some historians is questionable. In order to facilitate the tracing of households in different censuses, Thernstrom eliminated any households with very common first and last names<sup>32</sup>. Though it is understandable why he would like to use such a procedure, it is not defensible statistically since the characteristics of individuals with very common names are not the same as those with uncommon names.

Another criticism of most, though not all, studies of urban areas is that they are not designed to investigate differences within those communities. Most of the new urban historians do not attempt to link individual level census data with their location within the city. A study of marital fertility at the household level in two Boston neighborhoods, however, has demonstrated the importance of not treating the city as a homogenous entity<sup>33</sup>. There are some urban historians who have tried to cope with this issue in their research. The two large-scale urban history projects that have dealt the most effectively with the problems of space and neighborhood are Theodore Hershberg's analysis of Philadelphia and Olivier Zunz's study of Detroit<sup>34</sup>. In both of these studies, there is a conscious effort made to study family life within the context of their local neighborhood within the city.

Since many of the issues relating to the quality of the individual level census data have already been touched on in the previous section, we will focus instead on some of the conceptual and empirical problems of using these data. One of the

Michigan, 1850–1880, in: *Journal of Social History*, 5 (1971), pp. 26–45.

<sup>32</sup> Thernstrom, *The Other Bostonians*.

<sup>33</sup> Hareven, Tamara K., and Vinovskis, Maris A., *Marital Fertility, Ethnicity, and Occupation in Urban Families: An Analysis of South Boston and the South End*, in: *Journal of Social History*, 9 (1975), pp. 69–93.

<sup>34</sup> For a detailed discussion of the Philadelphia Social History Project, see Hershberg, Theodore, *The Philadelphia Social History Project: A Methodological History*, Doctoral dissertation, Stanford University 1973; Hershberg, T., guest editor, *A Special Issue: The Philadelphia Social History Project*, in: *Historical Methods Newsletter*, 9 (March–June 1976). Zunz, Olivier, *Detroit en 1880: Espace et Ségrégation*, Center for Research on Social Organization Working Paper, No. 121, University of Michigan 1975.

most widely discussed and analyzed issues is the problem of measuring social mobility using the available occupational scale. The development of an occupational scale is one of the major problems in this area<sup>35</sup>. Nineteenth-century occupations are not identical to those in the twentieth century. For example, whereas school teachers have a relatively high status today, this was not true in the past. In fact, many female school teachers left the classroom for the factory which paid them much better wages in the ante-bellum period<sup>36</sup>. Though this change of jobs was not seen as downward mobility by most people in the nineteenth century, it would appear as downward mobility in most of the social mobility studies which rank occupations into professional, semiprofessional, white collar, skilled, semiskilled, and unskilled categories. Similarly, the status of shoemakers varies greatly over time since they were considered skilled artisans in Lynn in the 1830's and 1840's, but became more like factory workers after the Civil War with the introduction of new technology and organization in the shoe industry<sup>37</sup>.

Another problem in studies using individual level data is that the occupation of the head of the family is used as the index of its status and well-being without taking into consideration the number of other wage-earners and dependents on that family. Recently, we have tried to develop a broader approach to this issue<sup>38</sup>. Though the occupation of the parent is a very useful and important indicator of the economic situation of the family, it is not the only economic data we would like to have. Ideally, we would measure the actual consumption needs of the family, as several contemporary studies have done. Unfortunately, such data are unavailable to us historically. We can go beyond just the occupation of the head of the household, however, by taking into consideration the number of individuals in the family who are employed as well as the number of consumers within that family.

Since the earning and consuming ability of individuals varies by age and sex, we adjusted our data by a set of weights to take these factors into consideration. Our work/consumption index is therefore a crude measure of the number of working units in each family divided by the number of consuming units. Though this index does not fully capture the individual family variations in income and consumption

<sup>35</sup> Many articles point to the problems in classifying occupations. See for example, Griffen, Clyde, *Occupational Mobility in Nineteenth Century America: Problems and Possibilities*, in: *Journal of Social History*, 2 (1972), pp. 310-330; Katz, Michael, *Occupational Classification in History*, in: *Journal of Interdisciplinary History*, 3 (1972), pp. 63-88; Conk, Margo Anderson, *Occupational Classification in the United States Census: 1870-1940*, in: *Journal of Interdisciplinary History*, 9, No. 1 (Summer 1978), pp. 111-130.

<sup>36</sup> Bernard, Richard M., and Vinovskis, Maris A., *The Female School Teacher in Ante-Bellum Massachusetts*, in: *Journal of Social History*, 10, No. 3 (Spring 1977), pp. 332-345.

<sup>37</sup> Dawley, Alan, *Class and Community: The Industrial Revolution in Lyon*, Cambridge/Mass. 1976.

<sup>38</sup> Kaestle, Carl F., and Vinovskis, Maris A., *From Fireside to Factory: School Entry and School Leaving in Nineteenth-Century Massachusetts*, in: Hareven, Tamara K. (ed.), *Family Transitions and the Life Course*, New York 1978; Mason, Karen, et al., *Determinants of Women's Labor Force Participation in Late Nineteenth-Century America*, in: op. cit.

needs, it does provide at least a beginning toward measuring a family's economic situation rather than just relying on information on the head of the household.

In our analyses of children attending school or the participation of women in the labor force, we used the work/consumption index in conjunction with the occupation of the head of the household. Though we anticipate that our particular formulation of this index may eventually be modified as researchers experiment with different weighting systems, it has proven to be useful both conceptually and empirically in our studies to date.

These are only a few of the conceptual problems involved in using individual level census data. Though I shall not produce more examples of other types of new and useful indices that can be developed from the manuscript census, such as an index of marital fertility, it should be pointed out that historians have not been particularly imaginative or aggressive in developing new ways of using the census data at the individual level. Many of the recent efforts along these lines are the result of trying to imitate as best as possible from the available census manuscripts some of the more interesting indices that have been developed by other social scientists studying the contemporary family.

Perhaps the most important advance in the use of individual level census data is the effort to use these data to estimate life course patterns. That is, historians are now starting to use the age-specific data from the manuscript censuses to reconstruct the probable life course experiences of individuals in the past. This is a particularly fruitful endeavor when we have data from more than one census and can follow different age-cohorts over time<sup>39</sup>.

The effort to reconstruct the life course of individuals from census data is very difficult because we usually cannot follow the same individuals over time. Instead, historians have to recreate artificial cohorts of individuals based on age-specific rates of individuals in different time-periods. The problem is that the people living in Lynn in 1860 are not necessarily the same ones living there in 1880. If an area experiences considerable in- or out-migration, as most urban communities did, we may have a very distorted picture of a cohort's life experiences on the basis of cross-sectional data from a small geographic area<sup>40</sup>. One way of minimizing this problem (or at least of sensitizing ourselves to it) is to use the aggregate age-specific census returns in conjunction with age-specific mortality estimates to calculate the net migration in the communities which we are investigating. In this way we can at least hazard some guesses about the type of biases introduced in our analysis by the fact that we have not drawn our samples from a closed population<sup>41</sup>.

<sup>39</sup> Elder, Glen H., *Age Differentiation and the Life Course*, in: *Annual Review of Sociology*, 1 (1975), pp. 165–190; Vinovskis, Maris A., *From Household Size to the Life Course: Some Observations on Recent Trends in Family History*, in: *American Behavioral Scientist*, 21, No. 2 (November/December 1977), pp. 263–287.

<sup>40</sup> Wells, Robert V., *On the Dangers of Constructing Artificial Cohorts in Times of Rapid Social Change*, in: *Journal of Interdisciplinary History*, 9, No. 1 (Summer 1978), pp. 103–110.

<sup>41</sup> Kaestle, *From Fireside to Factory*; Vinovskis, *From Household Size to the Life Course*.

Very few of the historical studies using the manuscript censuses have tried to analyze the life course of individuals. Despite the inevitable, and in some ways insoluble, methodological problems associated with such an approach, it is a much better way of organizing and operationalizing our data. A particularly useful perspective on the life course approach is provided in the writings of Glen Elder — a sociologist who has dealt extensively with family patterns and changes historically using a life course approach<sup>42</sup>.

Finally, I will close by considering some of the statistical techniques that have been used to analyze the individual level census data. Most historians have cross-tabulated their data with the inevitable and obvious shortcomings of such a procedure. A few have even introduced the use of multiple regression analysis with dummy variables to deal with categorical variables available for individuals from the manuscript census returns.

In our own work with the individual level census data, we have found that neither cross-tabulation nor multiple regression analysis suits our needs. The cross-tabulation of data simply cannot handle the complexity of factors we want to control in our analysis. Multiple regression analysis using dummy variables is quite adequate from a statistical perspective, but it is very unsatisfactory in terms of presenting the results to other historians who are less mathematically oriented. Therefore, we have turned to multiple classification analysis (MCA) instead<sup>43</sup>.

Multiple classification analysis is a form of multiple regression analysis with dummy variables which express results in terms of adjusted deviations from the grand mean (overall average) of the dependent variable of each of the various classes of the predictor variables. For example, MCA answers the question: how much of the likelihood of going to school was associated with being the child of an unskilled laborer, while controlling for other variables such as the age of the child, the ethnicity of the parents, and the community in which the child lived? Similarly, it provides an approximate answer to the question: *ceteris paribus*, what is the effect on youths' school attendance of the family's life course stage as measured by the age of the parents? MCA „controls“ for other variables by assuming while it looks at one class of a predictor variable that the distribution of all other predictor variables will be the same in that class in the total population, thus „holding constant“ their effects. Although traditional multiple regression programs also do this, MCA has three advantages: it does not require variables to be interval variables, it does not require or assume linearity and thus can capture discontinuities in the direction of the association and, finally, it is useful descriptively because it presents the reader with the gross effects of a predictor class, that is, the actual mean of each class, as well as the mean after adjusting for the influence of other variables. As historians learn to use MCA in their analyses of the individual level census data, they will find it to be a very useful way of analyzing categorical data as well as a relatively simple way of presenting their complex findings to our colleagues.

<sup>42</sup> Elder, Glen H., *Children of the Great Depression: Social Change in Life Experience*, Chicago, 1974.

<sup>43</sup> Kaestle, *From Fireside to Factory*; Mason, *Determinants*; Rosenkrantz, *Invisible Lunatics*.

## Reconstructing Biological Frameworks of Populations in the Past

„Since World War II . . . whole new fields, such as historical demography, and entirely new techniques, such as computer data processing, have appeared. . . . Historians have begun to raise questions previously unasked and to undertake research that once was thought impossible“.

Robert I. Rotberg and Theodore K. Rabb, editors of *The Journal of Interdisciplinary History*, presenting the new journal in the first issue, Fall 1970, Vol. 1, Nr. 1, p. 3.

The following research note might serve as a modest illustration of this quotation. Between 1975 and 1978 the History Department of the Free University of West Berlin, with the financial support of the Deutsche Forschungsgemeinschaft and the Volkswagen Foundation, has established a data bank, the core of which consists of around 7,000 reconstituted families from eight neighboring Hessian villages (the so-called Schwalm region) from the 16th to the 20th century. For complementary analyses a number of further sources were taken into consideration, e. g. parish registers, or aggregated demographic materials (vital rates, censuses) for the city of Berlin or the village of Gabelbach, west of Augsburg in Bavaria<sup>1</sup>. The source material consulted, the quantitative methods applied for its processing, the thematic goals of the project, as well as first concrete results have been described extensively in a number of papers published in 1976 and 1977, so that in the framework of this paper we can omit such issues as collection and selection of the material, critique of the sources, methods of data processing, etc.<sup>2</sup>.

In this note I would like to present some material from research work in progress and formulate a few thoughts, concentrating on one single point. In conscious reference to the title of the recent book by the British social anthropologist, Alan Macfarlane, *Reconstructing Historical Communities*<sup>3</sup>, this single point can be de-

<sup>1</sup> I wish to express my gratitude to the Deutsche Forschungsgemeinschaft, Bonn/Bad Godesberg, and to the Volkswagen Foundation, Hannover, for their generous financial support in carrying out this research.

<sup>2</sup> Cf. the author's: *Généalogie et démographie historique en Allemagne*, in: *Annales de Démographie Historique* (1976), pp. 77–108; *Historical Demography as Social History: Possibilities in Germany*, in: *Journal of Family History*, 2 (1977), pp. 305–332; *Historical Demography in Germany: A Research Note*, in: *Historical Methods Newsletter*, 10 (1977), pp. 122–126; *Mortalität in Berlin vom 18. bis 20. Jahrhundert*, in: *Berliner Statistik*, 31 (1977), pp. 138–145.

<sup>3</sup> Macfarlane, Alan, in collaboration with Harrison, Sarah, and Jardine, Charles, *Reconstructing Historical Communities*, Cambridge 1977.

scribed as „reconstructing biological frameworks of populations in the past“. Difficulties with the reconstruction of the biological framing conditions — which in my opinion, must constitute just as essential a part of Macfarlane's „reconstructing historical communities“ as, for instance, the reconstruction of marital areas, administrative units, economic ties, ritual areas, etc. — come from the dilemma in which we continually find ourselves: the individual elements of which the biological frameworks are composed, e. g. fertility, fecundity, fecundability, masculinity, live births and death rates, stillbirths, infant mortality, life expectancy, etc., are not autonomous essentialities. They are all more or less two-faced. They are only partly biological phenomena; they are also stamped by social, economic, intellectual, culture-geographic, and other framing conditions or factors. One can never say with certainty, where the biological sphere ends and where the influence of the just mentioned conditions or factors begins. One should not be led astray by the numerous print-outs and plotted graphs of the computer with their exact percentage values and precise curves. The tension which results from this dilemma can be very fruitful for research; it can, however — because of the numerous problems, most of which can be solved only in inter-disciplinary cooperation — just as easily lead to premature resignation or to an one-sidedness of interpretation, which is no longer adequate to the respective topic.

Our first concrete example is the monthly distribution of the number of conceptions — reckoned backwards as nine months before birth. Here the question can be posed, whether, or to what extent, the often observed decline in the number of conceptions (in Catholic areas at the time of the Ancien Régime) in the month of March is due to human biology, or to the influence of the (Catholic) church, which during Lent may have urged a certain abstinence also in sexual matters upon the faithful<sup>4</sup>.

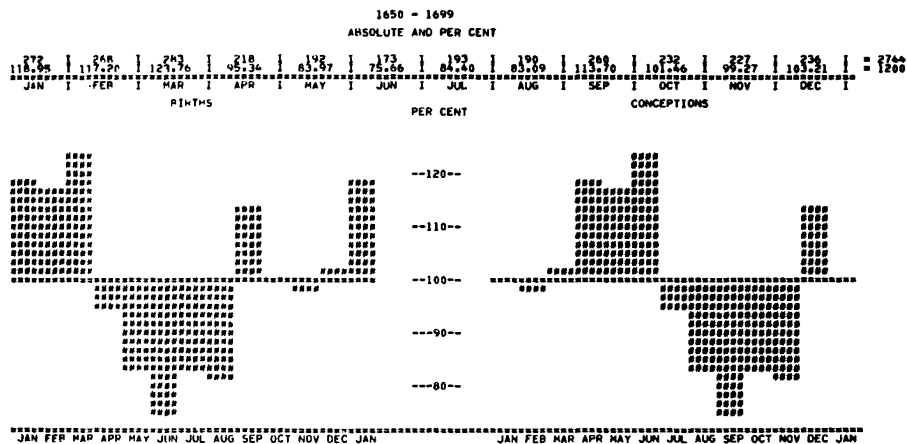
Or: how does one explain the likewise often demonstrated reduction in the number of conceptions (in agrarian societies of the European Ancien Régime) in mid and late summer as well as in autumn (August to November)? Is this primarily the effect of the greater intensity of labor in this season, which resulted in a decline in sexual activity in humans or is it part of a larger strategy of nature of yearly rhythms of sexual activity, the function of which might have been to bring as many children as possible into the world at a time when their chances of survival would have been the greatest. Doubtless, these chances were much better in the spring than in summer, on the one hand, because the mothers could better devote themselves to suckling children in winter and spring than in summer with its generally greater intensity of labor, and on the other hand, because a number of infectious

<sup>4</sup> Lebrun, François, *Demographie et Mentalités: Le mouvement des conceptions sous l'Ancien Régime*, in: *Annales de Démographie Historique* (1974), pp. 45–50; cf. also Smith, Daniel Scott, *A Homeostatic Demographic Regime: Patterns in West European Family Reconstitution Studies* in: Lee, Ronald D. (ed.), *Population Patterns in the Past*, New York 1977, p. 38.

diseases (especially e. g. diarrhea) in the summer regularly led to a greater number of infant deaths<sup>5</sup>.

Given a sufficiently large data bank, the computer can serve as an extremely useful instrument for research work in progress in the clarification of such questions. In line with the preliminary theoretical considerations touched on above, it can collate the distribution of the monthly number of births or conceptions for particular space and time units, or for particular sequences of time units, but also for particular social or occupational groups, and it can present them graphically for an — often stimulating — illustration (cf. Figure 1).

Figure 1: Monthly Distribution of the Number of Births and Conceptions in the Schwalm Region 1650 to 1699. (As presented graphically by the computer, the deviations are from the monthly mean value of 100; the sum for 12 month being 1,200).



A second example with a similarly posed question concerns the distribution of stillbirths over the months of the year (cf. Table 1).

The most interesting part of the table is doubtless the rubric, „Stillbirths as a percentage of total births“. It shows clearly the higher percentage of stillbirths in the winter months as compared to the summer months. The month of June represents the only exception in the rhythmic course of the year. For an interpretation here, it is necessary to weigh carefully a large number of factors, especially in view of the seasonally different labor-burden of women in agrarian societies in the household (different according to size and structure of the household, whether with or without servants), in the fields (e. g. for the hay harvest in June, but not for the

<sup>5</sup> Cf. the author's: Structures of Mortality in the 18th Century as Derived from Mass Statistical Analyses, in: Zeitschrift für Bevölkerungswissenschaft (1976), pp. 103–117.



*Table 1: Monthly Distribution of Stillbirths in the Schwalm Region 1570–1960*

Month	Stillbirths		Births absolute	Stillbirths as a percentage of total births
	absolute	percent		
January	51	10.63	2 175	2.34
February	42	8.75	2 083	2.02
March	42	8.75	2 170	1.94
April	37	7.71	1 924	1.92
May	32	6.67	1 740	1.84
June	34	7.08	1 601	2.12
July	29	6.04	1 925	1.51
August	34	7.08	1 940	1.75
September	35	7.29	2 037	1.72
October	41	8.54	1 899	2.16
November	47	9.79	1 858	2.53
December	56	11.67	2 001	2.80
Total	480	100.00	23 353	2.06

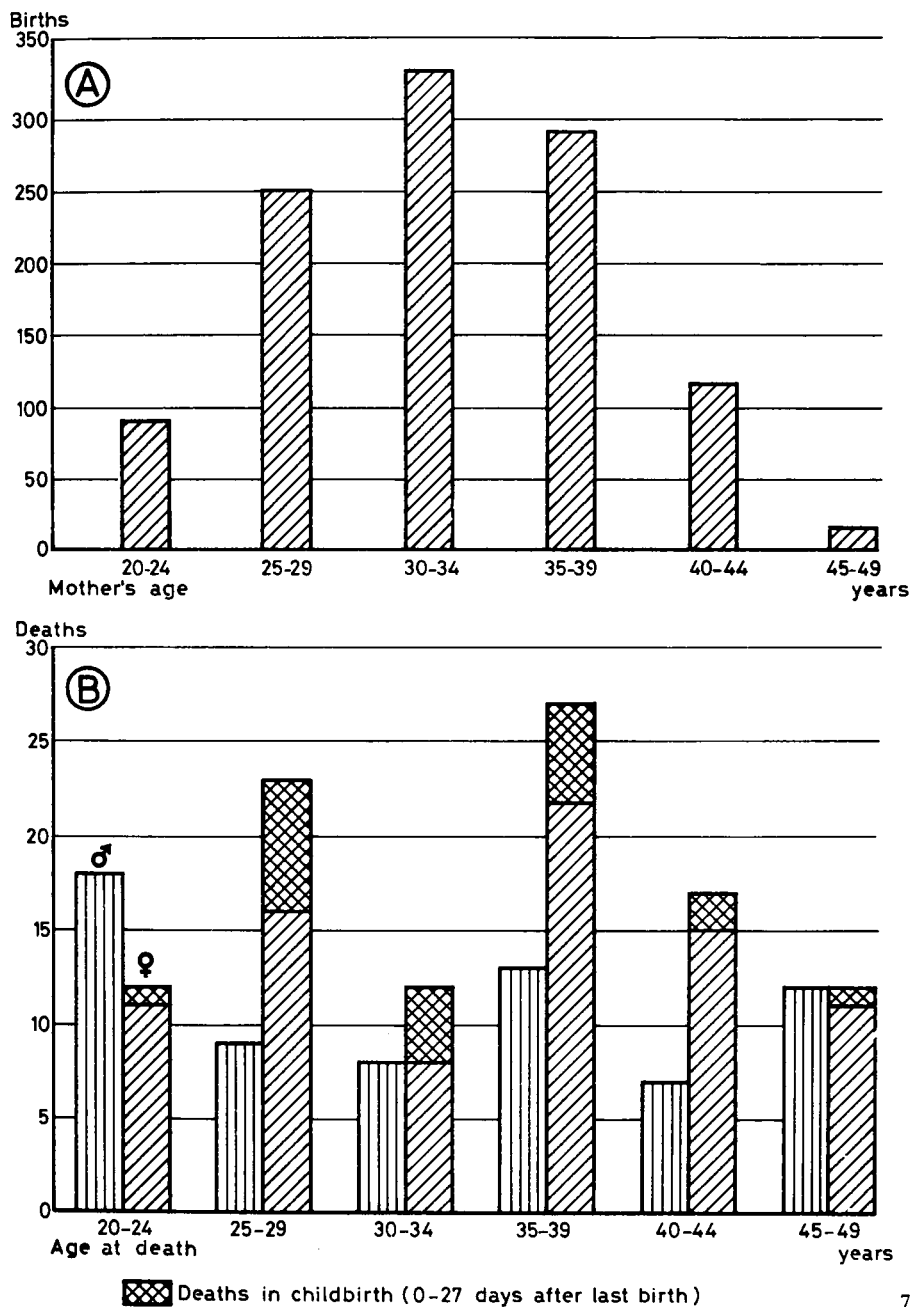
grain harvest in July and August), in the care of livestock. One must also take into consideration the seasonally conditioned differences in nourishment, in living and heating conditions, etc.<sup>6</sup>. If the data on stillbirths were great enough – in our material with 480 stillbirths all told for the entire period of time from 1570 to 1960, they are not – one could divide them into shorter time units. It might then be possible to ascertain, whether and when a change in the attitude of women, the family, or the society toward developing life took place, when and where more consideration was paid to pregnant woman, or when they themselves were better able to take care of themselves, and with greater probability to bring their children alive into the world<sup>7</sup>.

The example presented graphically in Figure 2 for the village of Gabelbach, 1680–1899 comes likewise from the border area between biology and women's living conditions.

<sup>6</sup> Cf. on this topic the two books by Léridon, Henri, *Aspects biométriques de la fécondité humaine*, Paris 1973, and *Natalité, Saisons et conjoncture économique*, Paris 1973, p. 76 (with an analogous monthly table for France for the years 1957–1959), p. 83.

<sup>7</sup> On the development of such parameters Edward Shorter bases an entire „New Agenda for psycho-medical history“, the subtitle of his article *Maternal sentiment and death in childbirth*, in: Branca, Patricia (ed.), *The Medicine Show. Patients, Physicians and the Perplexities of the Health Revolution in Modern Society*, New York 1977, pp. 67–88.

Figure 2: Excess Mortality of Women Between the Ages of 25 and 44 in Gabelbach, 1680–1899.



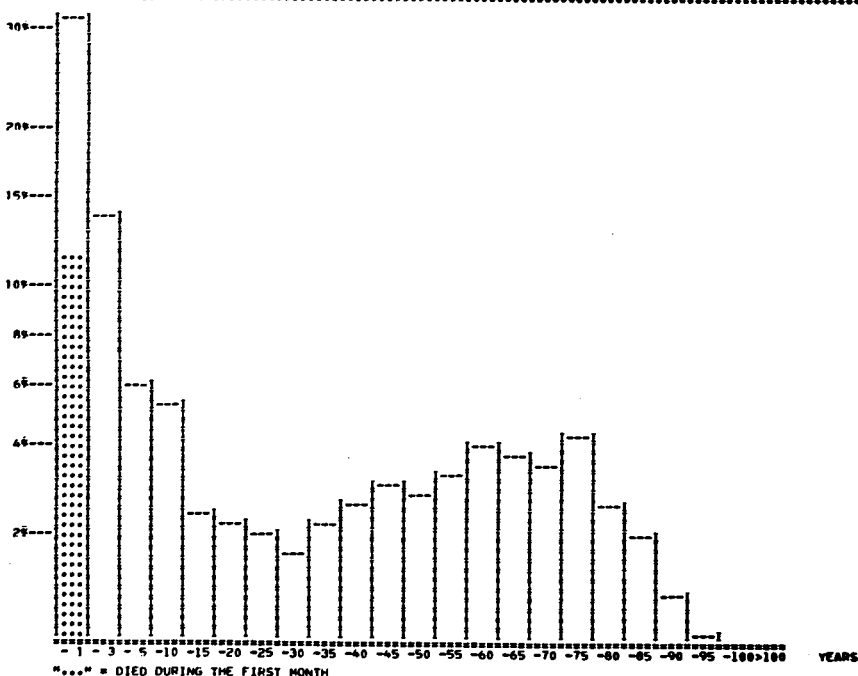
The diagramm shows, above, the number of children, which women in different age groups brought into the world. Below, the number of deaths of men and women in these same age groups is presented. It is clear that more women between the ages of 25 and 44 years died than men of the same age; but it is also easy to see from the figure, that this fact is due only in part to the death in childbirth of women of childbearing age. Evidently the inner-familial and -household position of the wife and mother, who was charged with the care of infants, is just as essential. In those days these infants very often died of various infections diseases — a situation which often also led to an impairment of the health of the mother who cared for

Figure 3: Number of Deaths at Different Ages as a Percentage of Total Deaths in the Schwalm Region, 1700–1749 and 1750–1799.

A

1700 - 1749

YEARS	-1	-3	-5	-10	-15	-20	-25	-30	-35	-40	-45	-50	-55	-60	-65	-70	-75	-80	-85	-90	-95	-100	>100	SUM
NUMER	662	296	176	114	50	44	40	34	47	50	44	60	71	85	81	76	94	57	36	17	3	0	1	2121
% MONTH	24.8																							
POP. CENT	31.7	14.5	4.4	5.4	2.4	2.6	2.4	2.1	2.7	3.1	3.5	3.3	3.4	4.5	4.3	4.1	4.4	3.2	2.3	1.3	.6	.5	.5	100

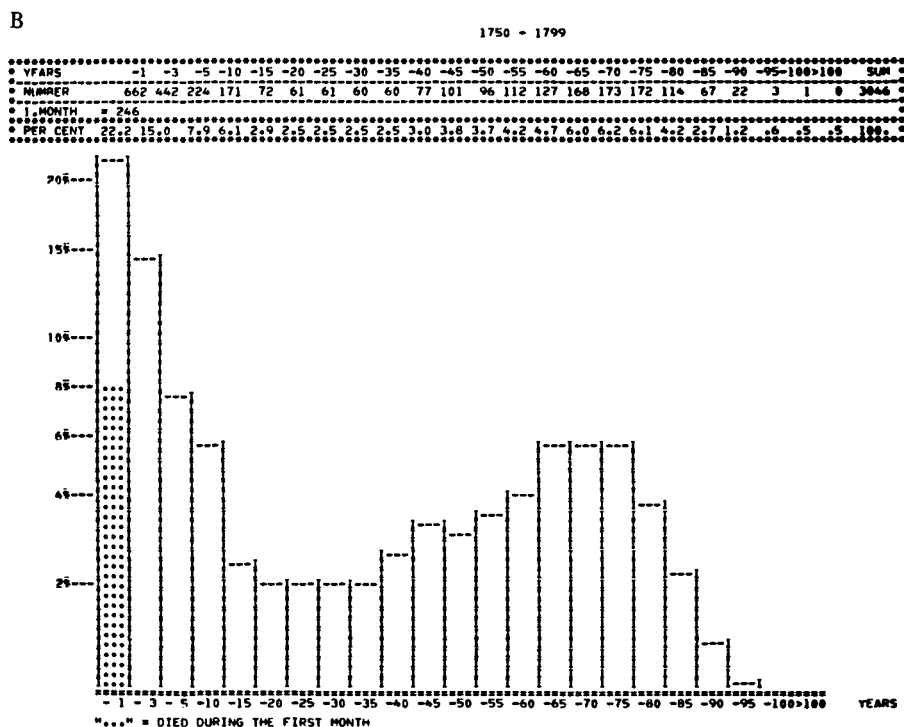


"..." = DIED DURING THE FIRST MONTH

them (whether through the additional labor-burden or through direct infection) and not seldom brought on her premature death<sup>8</sup>.

With this example, we have touched on a further central element of the biological framing conditions of human existence in general: the mortality patterns. One of the simplest methods to arrive at a more definite notion of the death patterns of a particular population at a particular time, consists in sorting the deceased according to their ages at death. For a graphic presentation, a logarithmic scale is to be recommended, since up to and even into the 20th century, mortality in the youngest age groups was incomparably greater than in all other age groups. In order to provide an orientation in the temporal succession of age specific death patterns in a somewhat simplified but very vivid and illustrative manner, the computer can be employed as a tool and can print out the corresponding diagrams in time-unit sequences (cf. Figure 3).

Figure 3: (continued)



<sup>8</sup> On the problem of masculinity at different ages (number of men per 100 women) cf. Delille, Gérard, Un problème de démographie historique. Hommes et femmes face à la mort, in: Mélanges d'archéologie et d'histoire de l'École Française de Rome, Paris 1974, pp. 419-443.

With the application of this method to our example from the Schwalm region and in the comparison of the two time periods, 1700–1749 and 1750–1799, one can clearly see a segment of that fundamental change in the circumstances of death in the course of the last two or three centuries: the gradual decline of the earlier, extremely high infant mortality (from 1700 to 1749 more than 30 % of all deaths were among infants under one year of age) and the successive increase of deaths in ever higher ages, which in the end has led to the phenomenon, well known today, of the superannuation of entire populations (e. g. in Western Europe).

On the basis of our data bank material however, the computer was also able to provide us with more precise accounts of the mortality framework of populations in the past in the form of life tables and to present them graphically in temporal succession. It thus shows how many individuals out of 1,000 born alive in a particular period were still alive at the age of 1, 2, 3, 4, 5, etc. respectively. We need not expound here on the immensely important role, which rising or sinking life expectancy at different ages plays for the economic, social and cultural development of a particular population. One need only think, for instance, of the increasing or decreasing problem of orphans and widows, of the influence on the age at marriage, of the varying quotas of life long celibacy, of increasing or decreasing population pressure (relation between population size and means of subsistence), of a greater or smaller immigration or emigration, as well as the consequences for the social, economic, and mental stability of the population.

On the basis of the life tables, it was possible to go even one step further. In connection with the study of Kjeldsen's on the „Evaluation of the impact of various diseases on mortality“<sup>9</sup>, we present two life tables for the Schwalm region for the period 1600–1649, one on the basis of all dead persons and a second time excluding those deaths due to the plague, which was recurrent in that period (cf. Table 2 and Figure 4).

Figure 4 shows clearly that the average life expectancy was drastically reduced on account of the plague epidemics in the 1610's and in the 1630's. Interestingly, this observation does not apply to infants and small children — at least not in our figure. Here we are probably dealing with a problem of source material rather than with biological reality. In historical demography the notorious underregistration of deaths among infants and young children in these earlier times is a well known fact. The Schwalm region is certainly no exception in this regard. Plague years, however, seem to be excepted from this rule, for in these years, infants and small children who died „of the pestilence“ are registered in the church books in great numbers.

<sup>9</sup> Kjeldsen, Kjeld, Evaluation of the impact of various diseases on mortality, in: Bulletin of the World Health Organization, 52 (1975), pp. 369–375. This article is based on Danish material for the years 1969 and 1971. Cf. also Preston, Samuel H., Demographic and Social Consequences of Various Causes of Death in the United States, in: Social Biology, 21 (1974), pp. 144–162.

Figure 4: Survivorship Proportions in the Schwalm Region, 1600–1649;  
a) on the basis of all registered deaths,  
b) with a specific registered cause of death (plague) excluded.

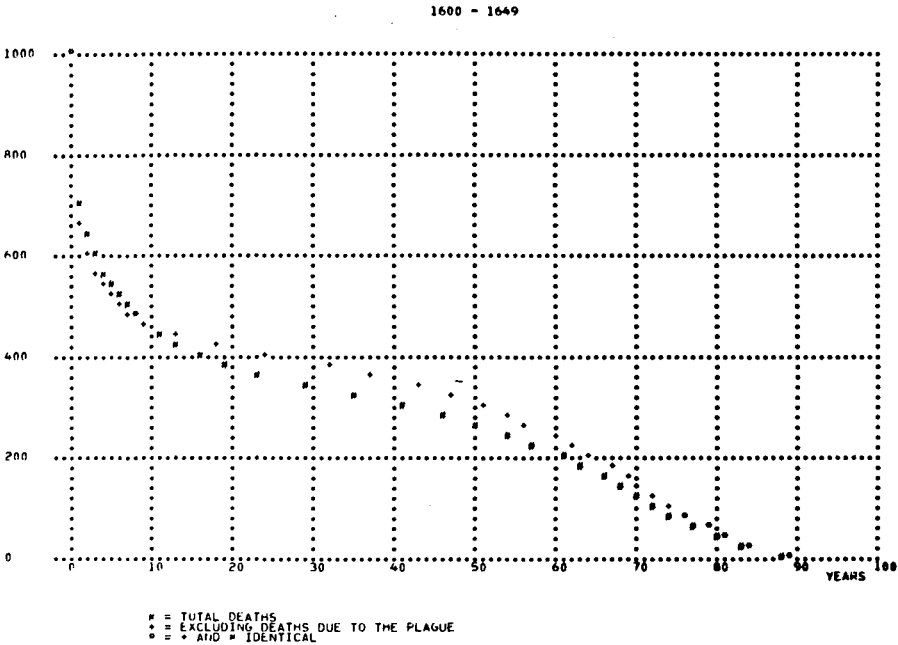
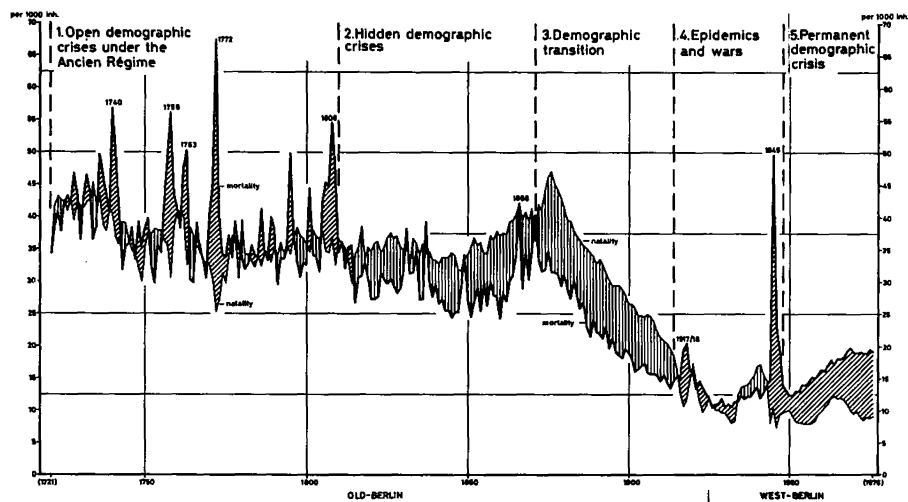


Table 2: Number of Deaths in the Schwalm Region, 1600–1649;  
a) total, b) due to the plague

Period	Deaths	
	a) total	b) due to the plague
1600–1609	64	1
1610–1619	356	113
1620–1629	341	41
1630–1639	497	142
1640–1649	253	0
1600–1649	1511	297

One final example from Berlin is still to be introduced. It shows the long range development of the crude birth and death rates per 1,000 inhabitants within the respective borders of the city<sup>10</sup> (cf. Figure 5).

*Figure 5: Births and Deaths per 1,000 Mean Population in Berlin 1721–1976.*



When one examines this graph, it is only natural at first to focus one's attention on the striking black peaks, which show a sudden increase in mortality up to double, triple or quadruple the normal level. The dates for the most important of these peaks are given in Figure 5: 1740, 1758, 1763, 1772, etc. up to 1945. It is in no case difficult to trace the respective increases in mortality to one or more of the three classic causes: war, famine, epidemic. The catastrophe at the end of World War II left its mark on the population of Berlin just as did the combination of the effects of war, hunger and pandemic influenza in 1917/18. Even where one's own memory or what one knows from hearsay are not sufficient, the sources necessary for the interpretation of such mortality maxima are abundantly available. Catastrophies of all kinds have always been amenable to the production of sources and make the historian's work easier. The high mortality of the early 1740's and 1770's is clearly related to bad harvests throughout Europe. During the Seven Years War (1756–1763) Berlin was occupied by the Austrians in 1757 and by the Russians in 1760; from 1806 to 1808 the city had to bear with a French occupation. Finally, 1866 represents the highpoint of a series of cholera waves, which began in the 1830's.

<sup>10</sup> For an exact description of the sources consult *Berliner Statistik*, 31 (1977), p. 145.

However historical research today does not so much inquire into such striking, individual events, but rather into longer range fundamental structural changes. For this purpose I have divided the entire era of 256 years observed here into five periods. The first period (from 1721 to approximately 1810) shows the last phase of a century-old model of mortality. In historical demography this is called „mortality of the type, Ancien Régime“. While the natality curve remains comparatively stable, the mortality curve shows considerable fluctuation from year to year. There occurred again and again open crises in which the number of deaths considerably exceeded the number of births.

A second phase shows itself in the years from 1811 to 1871. The range of the fluctuations in mortality was reduced to around a third of its previous extent, and the mortality curve remains throughout below the natality curve. Even mortality maxima such as that of 1866 no longer exceed natality. In the terminology of historical demography (French), it has become customary to designate such phenomena as *crises larvées*, i. e. hidden crises, which no longer resulted in a decrease in population.

A third period began in the 1870's and lasted until World War I. In the course of these four decades, mortality and natality declined from century-old average values of 30 or 40 per 1,000 mean population to approximately one third of their respective former values. This is not the place to go into detail about the causes of this demographic transition. As is well known, many different arguments are advanced as being of primary importance, for instance, the greater efficiency of state administration in securing the means of subsistence, long range improvements in the climate with a decrease in bad harvests, the increasing importance of medicine for the individual and the family, better physical hygiene, the decrease in infant mortality on account of lower birth rates, improvements in the area of drinking-water and sewage systems, innovations in medicine and medicaments, the development of a social welfare apparatus, and many others<sup>11</sup>.

The fourth period, the time between the first and second World Wars, brought back irregular values like those of more than a century before. Comparable death surpluses had before been experienced only under the Ancien Régime. On the other hand, something fundamentally new begins to be seen in the last phase, in the years from 1950 to the present. There has never before been a permanent crisis, such as that of the last quarter century, in which mortality, for years on end, is double the birth rate. Such a development leaves a much more enduring mark on a population than any short-term crisis of the old type — no matter how open or obvious. It should not be overlooked that we are dealing with the effects of a situation specific to Berlin and with a superannuation, which at this time is extreme. Nonetheless this

<sup>11</sup> For a discussion of the different arguments cf. Lee, Robert W., Primary Sector Output and Mortality Changes in Early XIXth Century Bavaria, in: Journal of European Economic History, 6 (1977), pp. 133–162. With regard to the historical decline of fertility in Europe cf. a series of monographs issued by the Office of Population Research at Princeton University. (For Germany: Knodel, John, The Decline of Fertility in Germany 1871–1939, Princeton/N. J. 1974).



development can be observed as a tendency both in the Federal Republic of Germany and in the German Democratic Republic, where likewise for the past several years birth deficits can be demonstrated. It is not famine, war or epidemics which threaten society today but rather — as Pierre Chaunu says — the *Peste blanche* (the White Plague, i. e. suicide) and the prevention of life in the form of family planning and abortion. Death need no longer rage periodically among humans, as it did for centuries, so that — as Malthus said — a too quickly growing population adapts itself to the more slowly growing means of subsistence; an excess population no longer comes into existence.

A consequence of the thorough-going demographic transition in the course of the last two or three centuries is a radical change in the age at death. As late as 1875 in Berlin, 60.7 % of all live-born children did not live past their fifth birthdays, and only 5.1 % lived to be older than 70; in 1975 on the other hand, the number of those who died at the age of 70 or more had risen to 68.1 % of all deaths, whereas the corresponding number of deaths of those under five years of age had sunk to 1.1 %. This fact alone is sufficient to show a fundamental change in the panorama of the causes of death. Infants and small children, in light of their short life-span, can for example never die of chronic illnesses to the same extent as people who have reached the biologically upper boundaries of life. Alongside the demographic transition, an epidemiologic transition has taken place<sup>12</sup>.

In itself no justification is needed for the fact, that in a discussion of mortality patterns in a research note on „reconstructing biological frameworks of populations in the past“, the *deaths* of human beings have relatively long been the focus of attention. Nevertheless I would like to avoid the appearance, that this might have been a goal in itself, that I might, so to speak, be pursuing a history without humans or that they interest me only after death has claimed them. While detailed investigations of death rates, life tables, the causes of death, etc. are necessary, they represent merely partial studies for a more encompassing history. And here other questions are of central importance, for instance: What does all this mean for the human being, the *living* human being? I mean by this not so much the mastering of the consequences of an essentially longer average life today, the social, economic and medical problems of an increasingly older society; rather I ask about the influence on our concepts, notions and attitudes. „Life-long“ means something totally different today than it did in the 18th century; the same applies to the marriage vow „until death do ye part“. Earlier there were more orphans; today more children of divorced parents. The century-old institution of marriage as a lasting tie attains a completely new temporal dimension, for which the longer-lived human is perhaps not yet ready. Generational conflicts can increase in a similar manner, simply because more generations live at the same time. It seems to me as though we have not at all yet gotten used to the longer lease on life allotted us. One century,

<sup>12</sup> On the epidemiologic transition cf. several articles by Omran, Abdel R., e. g. A Century of Epidemiologic Transition in the United States, in: Preventive Medicine, 6 (1977), pp. 30–51.

in the course of which the essential changes have taken place, has not been sufficient. Do we not act in many areas as if our life expectancy were on the average 40 or 50 years? One thinks of the often small commitment of the everyday man to the active provision for a healthy old age.

However, not only the time up to death has changed, death itself has changed. It no longer grips us so often in childhood, youth, in the flower of our years; it no longer tears us out of profession, responsibility, family obligations; it is no longer a mass phenomenon as it was in the times of the great epidemics. Today it comes furtively and individually, often as a release, an end to years of chronic pain, often seemingly suddenly after we have already died professionally, socially, familially.

We should today have more opportunity and time to adjust ourselves to death and to prepare ourselves for it, to consider it as the close of a long and fulfilled life. But parallel to the successful exertions to postpone death until ever older ages, another process has taken place, through which death — wherever possible — has been repressed: out of consciousness, out of the family, out of everyday life. Only in most recent years has a movement begun, which encompasses not only various branches of science but also the mass media, whose goal is to remove the taboo of the „forbidden death“. It is to be hoped, that this is not merely a new variant of repression — similar to that involved in the marketing of sexuality — but rather the expression of attempts, better than before, to use the chances which have been made available to us through the fundamental changes in mortality patterns outlined here as an essential part of the biological framework of human beings during the last two or three centuries<sup>13</sup>.

This last example may be taken as an illustration of what arouses my interest in reconstructing biological frameworks of populations in the past. It is not merely biology of man in history per se, not even working out and interpreting larger inter-relationships in this context, such as demographic or epidemiologic transitions; it is one step further: the intricate but stimulating field between biology and man, between birth, health, disease, life expectancy, death, and man's attitude toward birth, health, disease, life expectancy and death; his arrangements with biology, with his body's functions or dysfunctions, his changing reactions within changing biological frames.

<sup>13</sup> Once again here, many suggestions and impulses are derived from Ariès, Philippe, *Western Attitudes toward Death from the Middle Ages to the Present*, Baltimore 1974. In the meantime, Ariès has filled out his own framework with more research: *L'homme devant la mort*, Paris 1977, (642 pages!).

## Theodore Hershberg

### Interdisciplinary Research at the Philadelphia Social History Project: Analytic Goals, Data and Data Manipulation Strategies for the Study of the Nineteenth-Century Industrial City<sup>1</sup>

#### Introduction

The QUANTUM-SSHA Conference is part of a larger process characterized by twin shifts among scholars on a world-basis. Electronic data processing, while not necessarily a part of the new scholarship, served as a catalyst and the widespread availability of this technology today no doubt underlies and accelerates the changes taking place. Historians are moving beyond impressionistic approaches and evidence to rely upon processed-produced data and social science methods and analytic techniques, while social scientists are moving beyond correlation and contemporary cross-sectional analyses to use time-series data and historical perspective in their research. This happy state of affairs is long overdue and we hope that the momentum can be sustained.

Not only has the new and powerful computer technology facilitated these transformations, but it places before us the opportunity to create a genuinely new form of research that is at once interdisciplinary, collaborative and historical. I emphasize that we have before us a possibility, not a certainty. As the last fifteen years of research in the United States makes clear, the mere use of quantitative data and methods will not lead inexorably to the successful establishment of this new research form. To create a cooperative and cross-disciplinary research organization, we will have to work consciously and deliberately toward bridging the gaps that separate the disciplines. This requires the integration of the disparate theories, con-

<sup>1</sup> The Philadelphia Social History Project (PSHP) is directed by Theodore Hershberg, Associate Professor of History and Public Policy. PSHP gratefully acknowledges the support it receives from the following agencies of the U.S. federal government: Center for Studies of Metropolitan Problems, National Institute of Mental Health (MH 16621); Sociology Program, Division of Social Sciences, National Science Foundation (SOC 76-20069); Division of Research Grants, National Endowment for the Humanities (RO 32485-78-1612); Center for Population Research, National Institute of Club Health and Human Development (HD 17413).

cepts, methods, techniques and knowledge created by the separate disciplines that bear upon common subject matter, but which over the years have been artificially segregated from each other.

The difficulty of this pioneering task is exceeded only by its importance. In meeting the challenge that it poses, we must keep clear in our minds that research ends and means are not unrelated, that the findings are themselves very much dependent upon the manner in which research is organized. While the experience of the Philadelphia Social History Project (PSHP) over the last nine years can be offered now as a successful model for emulation where the collection and processing of data and cooperation of scholars from different disciplinary backgrounds are concerned, it is still too early in our data analysis stage to claim that we have also mastered the many aspects of genuine interdisciplinary research. It is not clear that we have even identified, let alone resolved, the problems of organization and interpretation associated with such an approach to knowledge. Nevertheless, we have learned a good deal, perhaps enough to offer our experience as an encouraging though tentative beginning whose progress should be watched. It is our hope to share the details of this experience so that our mistakes can be avoided and our success learned from. Researchers associated with the PSHP are identified in Appendix I, a bibliography of PSHP research appears in Appendix II, and the machine-readable data base that supports our research is described in Appendix III.

The version of this paper presented to the Cologne Conference included a section, „Interdisciplinary Historical Research in the United States: Problems and Possibilities.“ This section has been removed from the current essay for the sake of brevity; however, readers can find a fuller discussion of these issues elsewhere<sup>2</sup>. In this paper, I will summarize my views on the importance of an interdisciplinary approach for urban historical research and discuss how research at the PSHP is organized: our data base, whose scope and detail makes possible the interaction of scholars from separate disciplines; several of the analyses such a data base sustains; and a variety of data manipulation techniques that make good on the quip that „the whole is greater than the sum of its parts“.

<sup>2</sup> See the introductory chapter of Hershberg, Theodore (ed.), *Toward an Interdisciplinary History of the City: Work, Space, Family and Group Experience in Nineteenth-Century Philadelphia*, Oxford University Press (forthcoming 1979–1980). This volume contains a selection of PSHP papers that report substantive findings from the first round of research. The origins and methodological development of the PSHP are detailed in Hershberg, Theodore, *The Philadelphia Social History Project: A Methodological History*, Doctoral dissertation, Stanford University 1973; and Hershberg, T., guest editor, *A Special Issue: The Philadelphia Social History Project*, in: *Historical Methods Newsletter*, 9, No. 2–3 (March–June 1976).

## The Organization of Research: The Need for an Interdisciplinary Approach

The argument offered is simple: a fundamental relationship exists between how knowledge is sought and the nature of the learning obtained. The ways in which we organize to conduct our research have a necessarily profound impact on what we are likely to find. This relationship owes its power to the revolution in the sociology of knowledge that has characterized the past century.

Guided by theories, concepts and methods developed in increasingly distinct and specialized disciplines, scholars expanded the frontiers of knowledge with great rapidity. By the end of the nineteenth century, scholarship had been organized in the major disciplines around which most universities are still structured; each discipline organized nationally and internationally in professional associations and generated a professional community of discourse of its own. The „Renaissance Man“ may have remained the ideal for undergraduate education, but he had been discarded, by the twentieth century, as an unacceptable model for the new professional, specialized research scholar<sup>3</sup>.

This reorganization of knowledge produced extraordinary progress, of course, but it has not been an unmixed blessing. Indeed, in important respects, professional specialization has progressed to the point of diminishing societal returns. The staggering amount of knowledge generated can barely be digested within a given isolated field, much less be made visible to others. As philosopher Stuart Hampshire recently pointed out, discussing the consequences of proliferating but unintegrated and undigested knowledge, „if we do not know what we know, then our first order knowledge is apt to be unused, almost if it did not exist“<sup>4</sup>. Even more seriously, disciplinary boundaries have tended to become barriers preventing us from seeing the contours of enormously complex problems in a real world that resists compartmentalization. Interdisciplinary effort can thus become not simply desirable for maximizing integrated knowledge, but an absolute necessity if problems are to be correctly understood and engaged.

This problem of fragmented knowledge and narrow disciplinary boundaries has long been recognized by students of the city; indeed, urban problems are perhaps the most frequently cited illustration of the more general necessity of interdisciplinary study. „The city is too complex a world to be understood from a single vantage point alone . . .“, noted Richard Wohl in a characteristic statement. „The full dis-

<sup>3</sup> Bledstein, Burton, *The Culture of Professionalism*, New York 1976; Kuklick, Henrika, *The Organization of Social Science in the United States*, in: *American Quarterly*, 23 (Spring 1976), pp. 124–141; Haskell, Thomas L., *The Emergence of Professional Social Science: The American Social Science Association and the Nineteenth-Century Crisis of Authority*, Urbana 1976.

<sup>4</sup> Hampshire, Stuart, *The Future of Knowledge*, in: *New York Review of Books*, 24 (March 31, 1977), p. 14.

covery of what has passed in a city's history can only be called forth by cooperative, interdisciplinary inquiry." „To study the city“, Allen Davis argued, „means to cut across the artificial lines of departments and disciplines“. And Theodore A. Brown concluded with some frustration that „there is no sense in which I can pursue studies in city history as a practitioner of a self-sustained ‚discipline‘ called history . . . if the approach is worth anything at all it needs a great deal of reinforcing from other so-called disciplines“<sup>5</sup>. Virtually every survey of urban history in the last two decades has come to precisely the same conclusion<sup>6</sup>.

But given this unanimity, the record of actual achievement in interdisciplinary work is disheartening. Social scientists and historians have been unable to generate

<sup>5</sup> Wohl, R. Richard, *Urbanism, Urbanity, and the Historian*, in: *University of Kansas City Review*, 22 (Autumn 1955), p. 57; Davis, A. F., *The American Historian vs. The City*, in: *Social Studies*, 56 (1965), p. 134; Wohl, R. Richard, and Brown, A. Theodore, *The Usable Past: A Study of Historical Traditions in Kansas City*, in: *Huntington Library Quarterly*, 23 (May 1960), pp. 237–259.

<sup>6</sup> No useful purpose can be served by presenting further testimony. Suffice it to say that all surveys of the urban history field, most notably those by Eric Lampard, Charles Glaab, Charles Tilly and Roy Lubove have made the same point. Lampard, Eric E., *American Historians and the Study of Urbanization*, in: *American Historical Review*, 67 (October 1961), pp. 49–61; Glaab, Charles N., *The Historian and the American City: A Bibliographic Survey*, in: Hauser, Philip M., and Schnore, Leo F. (eds.), *The Study of Urbanization*, New York 1965, pp. 53–80; Tilly, Charles, *The State of Urbanization*, in: *Comparative Studies in Society and History*, 10 (October 1967), pp. 100–113; Lubove, Roy, *The Urbanization Process: An Approach To Historical Research*, in: *Journal of The American Institute of Planners*, 33 (January 1967), pp. 33–39; at least at the level of rhetoric, such sentiments are shared by scholars in disciplines other than history. For example, one of the four major recommendations made by the Social Science Research Council's Committee on Urbanization was „to work towards the advancement of multidisciplinary research in the process of urbanization“. Hauser, Philip M., *Urbanization: An Overview*, in: Hauser, *The Study of Urbanization*, p. 41.

Sentiments of this nature were voiced not only by American urban historians, but by their European colleagues as well. *The Study of Urban History*, New York 1968, a widely respected collection of essays by scholars in Great Britain edited by A. J. Dyos, was filled with calls for cooperation across disciplinary lines; similar appeals regularly appear in the French journal *Annales*; and scholars in the Scandinavian countries have also supported the idea of interdisciplinary research. See also a special edition of *Annales* devoted to research on the city (1970), especially the remarks of O. Zunz; for work in Norway, see Langholm, Sivert, *The Ullenshaker and Kristiania Projects at the University of Oslo*, *QUANTUM-SSHA Conference* (Cologne, August 10–12, 1977); for work in Sweden, see Agren, Kurt, et al., *Aristocrats, Farmers, Proletarians: Essays in Swedish Demographic History*, Uppsala 1973; also see Öhngren, Bo, *Folk: Rorelse: Samhallsutveckling, Flyttnings-monster Och Folkrorelser: Eskilstuna, 1870 to 1900*, Uppsala 1974. For earlier work in Sweden, see the results of the conference organized by Carl Goran Andrae (Uppsala June 1973) and reported in a special issue, *History and the Computer*, in: *Historical Methods Newsletter*, 7, No. 3 (June 1974), p. 3; an extraordinary effort with enormous analytic potential is now underway to make machine-readable masses of historical information covering the development of Stockholm over a 300 year period; see Gustafson, Uno, *Demographic Data Base Stockholm*, Stockholm Stadsarkiv 1977.

new modes of research organization, whatever their commitment to the ideal of a new interdisciplinary scholarship. Consider the inadequacies of the usual responses to the call for innovation. There is first of all the interdisciplinary conference. Useful exchanges result and anthologies are often produced, but subsequent research remains fragmented. Second is the response of the individual scholar who strives to absorb the relevant literature of ancillary disciplines. The noble intention is one inevitably doomed owing to the limitations of what any individual can accomplish. The dilemma is a profound one, for as Lawrence Stone has observed, „to ignore the contributions of the social sciences is clearly fatal; to master them all, or even any one, is clearly impossible“<sup>7</sup>. Finally, in recent years, we have seen a proliferation of interdisciplinary studies programs (black, women, ethnic, etc.) and pan-professional associations. But undergraduate course sequences are not capable of sustaining active cross-discipline research, and pan-professional organizations (the Social Science History Association is the most recent example) measure and stimulate interest rather than implement such efforts. All of these responses, in short, facilitate consumption, not production of new approaches. They do not provide mechanisms for truly integrating the disparate research they encourage. They sustain interdisciplinary communication, but not interdisciplinary process.

The failure to accomplish more than this despite laudable intentions is not accidental. It is traceable, rather, to a combination of *structural factors* — the system of hiring and rewards governing the organization of research in universities — and *cultural factors* — that set of values sanctifying individual accomplishment while deemphasizing the collaborative efforts necessary to form the organizational base for interdisciplinary research<sup>8</sup>. Whereas researchers in the „hard“ sciences are accustomed to working collaboratively across disciplines, and assume the importance of systematically expanding knowledge built on the base of previous research, in the humanities emphasis has been on originality and creativity, often purchased at the expense of cumulative scholarly advancement.

The challenge is thus to develop mechanisms capable of transcending both structural and cultural obstacles in the humanities and social sciences, mechanisms that will have to work, moreover, without the integrating power that precise applied research has developed in the hard and policy sciences. An integrating mechanism of a different sort must be found.

Urban research remains the best ground on which to construct such a mechanism. As A. J. Dyos has pointed out, „There is a widening awareness of the possibility . . . that in the study of the way in which urban society has organized itself spatially and structurally and has developed such a variety of social systems, there exists an incomparable arena in which to bring together more explicitly a number of these

<sup>7</sup> Stone, Lawrence, *History and the Social Sciences in the Twentieth Century*, in: Delzell, Charles E. (ed.), *The Future of History: Essays in the Vanderbilt University Centennial Symposium*, Nashville/Tenn. 1977, p. 19.

<sup>8</sup> See Hershberg, Theodore, *The Organization of Historical Research*, in: AHA Newsletter, (October 1976).

converging disciplines“. Such a focus „lends itself so well to fruitful exchanges among the disciplines“, O. D. Duncan and Leo Schnore accurately observed, „because of its strong empirical base and its relatively concrete view of society“<sup>9</sup>.

The experience of the PSHP suggests that it is possible to construct and operationalize the needed mechanism. A common data base that includes both ecological and behavioral information about the city serves as the necessary facilitating device, and computer technology and the methodologies developed to manipulate the data can support the efforts of scholars trained in separate disciplines to integrate their knowledge about the city. Such an approach provides a working model of an interdisciplinary structure, and, even more, a collaborative and interdisciplinary research culture. To be sure, formidable obstacles face those who would emulate this approach, and the current financial crisis makes it clear that solutions may have to be found in crossinstitutional and perhaps regional and national cooperation as well. The technology of electronic data processing will support these efforts.

Historians studying the sociology of knowledge understand the relationship between organization and scholarship; it is time to apply such a critique to the present structure of research. To ignore it further dooms us to continued frustration. Unless new institutional mechanisms are developed to support interdisciplinary collaborative research, our knowledge about cities will remain hopelessly fragmented.

## A Common Data Base

Unlike interdisciplinary conferences or pan-professional associations that facilitate important discussion and exchange of ideas, a common data base actively shared makes it possible for researchers to implement in actual research activities the expertise developed in their respective disciplines. Beyond shared working and office space, prolonged use of the same data and criss-crossing perspectives on many of the same questions operate to break down barriers to communication. Daily contact and discussion ensure that disciplinary jargon gives way to a vocabulary that can be understood by all researchers. While common questions, common problems, and

<sup>9</sup> Dyos, H. J., *Agenda For Urban Historians*, in: Dyos, *The Study of Urban History*, p. 4; Duncan, Otis D., and Schnore, Leo, *Cultural Behavioral and Ecological Perspectives in the Study of Social Organization*, in: *American Journal of Sociology*, 65 (September 1959), p. 145. Different disciplines assign different meanings to the term „behavior“. As used by sociologists Duncan and Schnore, „behavior“ refers to the realm of the attitudinal; as used by social historians (and in PSHP research) „behavior“ refers to human actions – to what people did. Studies of behavior by social historians thus include fertility, marriage, occupational and residential mobility, mortality, etc.



even a common research site can *bring people together*, a common data base makes possible the integration of their efforts. The alternating use of the same data as dependent and independent variables, the ability to conduct research using common categorization schema and specific measure values, and the opportunity to formulate hypotheses drawing upon information normally outside the standing disciplinary paradigm — all of these operate in support of successful interdisciplinary collaboration. It is not that these things *cannot* be undertaken in research unsupported by a common data base, but that they rarely occur because of so many intellectual and logistical barriers.

The machine-readable data base of the PSHP consists of the following sources for Philadelphia County in the years 1850, 1860, 1870 and 1880. Each data set described below is further detailed in Appendix III.

(1) *Population Manuscript Schedules of the U. S. Census*: These schedules are by now a familiar part of the new social history in the United States. Our efforts commenced with the conversion of 100 rolls of microfilm to hard copy, an expense that has proved its value many times over. We now have 58,000 pages containing information describing each of the 2.5 million persons who lived within the county at the time of the four decennial enumerations (name, age, sex, race, occupation, property, place-of-birth, etc.). Extensive sub-sets of this information were converted to machine-readable form. Philadelphia was America's second largest city in these years with a population of 845,000 in 1880. When geographic location is added to these data, it is possible to examine a wide array of socioeconomic, demographic and spatial topics.

(2) *Manufacturing Manuscript Schedules of the U. S. Census*: Not as well-known as the population manuscripts, the information contained in these documents — product type, capital investment, wage levels, number and sex of employees, type of mechanization, and the kinds, quantities and values of raw materials and finished products — makes possible the detailed study of the city's industrial base and industrial geography as they underwent significant structural transformation. All the information in the schedules describing 29,000 *individual* firms were converted to machine-readable form. Throughout the period, Philadelphia's economy was highly diversified, it had a significantly disproportionate share of total U. S. manufacturing output, and it was among the world's leading centers of industry.

(3) *Business Directories*: Published on a yearly basis in most American cities throughout most of the nineteenth century and into the twentieth century, the business directories listed alphabetically by products sold or services offered the names and street addresses of all subscribers. Roughly 140 000 firms advertised in Philadelphia's directories in the four years indicated above. These data permit a wide variety of uses. They provide spatial information otherwise unavailable; they can be used to examine the degree of turnover across business types and areas of the city; they tell us some vital things about the non manufacturing sectors of the economy; and they constitute the best single source of information to reconstruct the range of shops, services and institutions throughout the city and are hence invaluable for micro-level study of neighborhoods or given areas of the city.

(4) *Transportation Facilities*: The exact routes followed by all horse-drawn street cars as well as passenger and freight railroads were converted to machine-readable form. The horse-car system emerged in the late 1850's when iron rail was laid in the streets of the city; because of significant friction reduction, three-to-four times as many passengers could be hauled by the horse-cars than was possible with the urban stagecoach which operated over cobblestone and dirt roads. By 1880, some 99 million passengers rode across the city's 300 miles of track. The system was not replaced with electrified trolleys until the mid-1890's. These data play an important part in the overall research, enabling us to break into the „chicken and egg“ relationships between transportation and the city's residential, industrial and commercial expansion. Transportation led the way in all of these and played a vital role in the larger process of the differentiation of urban space.

(5) *Public and Private Institutions and Associations*: Collected for the same years as the decennial census surveys and usually found in city directories, newspapers and other listings, the names and addresses of these public and private, religious and secular institutions and voluntary organization meeting sites were converted to machine-readable form. Along with other information described above, they permit the reconstruction of the urban environment in rich detail and enable us to differentiate among the wide variety of neighborhoods and other specialized sections of the city.

(6) *Death Registers*: The 19,000 deaths that occurred between July 1, 1880 and June 30, 1881, were converted to machine-readable form. They contain valuable information describing the deceased (name, age, sex, race, marital status, occupation, birthplace, address, date and cause of death, attending physician, place of burial) (Appendix III). These dates were selected because they followed immediately upon the enumeration of the population census in June, 1880. The information contained in these documents enables us to study how death brought on by different causes varied not simply across age and sex, but according to ethnicity, occupation, neighborhood and seasonality. When linked back to the population schedules, we can arrive at estimates of accuracy in the respective documents and make important adjustments in the base population figures necessary to calculate death rates; the linkage also makes possible the examination of the household and family context just prior to death. When associated with all other PSHP data, the different kinds of death can be correlated with the full range of ecological variables<sup>10</sup>.

(7) *Areal Units as a Common Denominator*: All PSHP data have been coded for geographic location within the 130 square miles of Philadelphia County. With an area so large, it becomes possible to conceptualize the city as a metropolitan region, containing the City of Philadelphia and many outlying communities all found within the county. The spatial system devised is essentially a grid network emerging from a series of vertical and horizontal lines drawn at fixed intervals over the map

<sup>10</sup> The research on mortality will be supported by a grant from the Center for Population Research, National Institute of Child Health and Human Development, Theodore Hershberg and Gretchen Condran co-principal investigators.

of the county. Of a total of roughly 7,000 grid units, each 660 by 775 feet, some 1,000 were inhabited in 1850 and 2,000 in 1880. There is no „ideal“ areal unit – different analyses require different size units. In many instances, we aggregate up from the grid level which reduces the number of areal units from several thousand to several hundred. The PSHP grid unit – small, unchanging over time and not drawn by politicians – has sustained a wide variety of analyses. As will be elaborated below, the use of common spatial codes makes it possible (through a variety of data manipulation techniques) to create a series of machine-readable records containing an infinite combination of variables with which to study simultaneously and systematically the relationships between individual-level behavior and ecological setting.

## The Evolution of Analytical Goals

Today it is possible to describe the central concern of the PSHP as the deepening of our understanding of the micro-level causes and consequences of urbanization and industrialization. Our earliest efforts were far less broad and their evolution was inextricably bound up in the expansion of the data base and the involvement of scholars from disciplines other than history. The research began in 1969 as a study of comparative social mobility among blacks, Irish and German immigrants who lived in nineteenth-century Philadelphia. The initial data collected were drawn from only the population manuscript schedules and described the personal attributes of the masses of ordinary people. Yet, despite the richness of the information made machine-readable and the power of computer technology to manipulate and assist in the analysis of these data, it became obvious that further information was required if the larger urban-industrial context in which these groups of people lived was to play an active explanatory part in the research. Considered alone, the population data confine the „city“ to a passive role – „urban as site“ rather than „urban as process“<sup>11</sup>.

The social mobility studies undertaken by American historians in the last decade or so provide a collective example of the „urban as site“ approach. Working alone with limited resources and the tools of a single discipline, they used data drawn

<sup>11</sup> For an elaboration of this distinction and a critique of the „New Urban History“ in the United States, see Hershberg, Theodore, *The New Urban History: Toward an Interdisciplinary History of the City*, in: *Journal of Urban History*, 5 (November 1978). For a Discussion of „Urban as Process“, see Hershberg, Theodore, et al., *A Tale of Three Cities: Blacks and Immigrants in Philadelphia, 1850–1880, 1930–1970*, for a special issue on Race and Residence in American Cities, in: *The Annals*, 441 (January 1979), pp. 55–81.

only from the population census schedules or city directories. They treated geographic and career mobility as their only dependent variables and used age, occupation, place of birth, property holding and other of the individual's personal attributes as their only independent or explanatory variables. The impact of urbanization and industrialization upon either the structure of employment opportunities or the rates or patterns of mobility, therefore, could not be directly or systematically assessed. Adding new data that described the city's industrial base and geography, its commercial and transportation facilities, and its public and private institutions made it possible to learn systematically how micro-level behavior and the urban environment interacted and how group experience was differentially affected. While only social historians were interested in the initial research, the addition of the new data attracted pre- and post-doctoral scholars from a variety of social science backgrounds. They represented economics, sociology, demography and geography and they brought with them the invaluable expertise for the study of the city developed in their respective disciplines. In other words, our experience followed this sequence: learning how the urban-industrial environment affected behavior required new explanatory variables; this necessitated the expansion of our data base; and once the data base expanded, we found that many social scientists wished to work along with us.

With input from the social scientists, the conceptual, methodological and technical tools became available that enabled us to treat the city in a dynamic fashion. The passive quality inherent in treating the city as an incidental setting for the study of behavior was replaced by a concrete view which saw the city as having particular form and substance that both affected and was affected by the people within it. Thus it is now possible to identify two, rather than only one, basic perspectives in our overall research design. Some PSHP researchers, drawn largely from the social sciences, treat the city as a dependent variable; they ask how the urban environment evolved, what it consisted of at given points in time, and how it changed under the impact of industrialization. A second group of PSHP researchers treat the city as an independent variable; they ask how aspects of the urban environment affected a wide range of behaviors, processes and attitudes. How did the city affect birth, marriage, family and death? What was the impact of the urban environment on migration, occupational and residential mobility, associational membership, social structure and the distribution of wealth? How were the processes of social mobility, assimilation, socialization and the development of political and class consciousness, values and personality affected by an urban-industrial setting?

From these distinct, but closely inter-related perspectives, four substantive areas of research have emerged: (1) the Nature of Work; (2) the Uses of Urban Space; (3) Developments in the Life Course; and (4) the Experience of Special Population Groups. The first of these explores the city's economy (particularly the composition and development of its industrial base), the composition of the labor force, the changing nature of the opportunity structure and the rates and patterns of career mobility. The second is concerned with the process by which the city was transformed from a geographically compact and heterogeneous mercantile center to a

sprawling and functionally distinct industrial city with growing divisions in land use for business and industry and segregation in residence along class, race and ethnic lines. The third area permits the large-scale phenomena with which we are concerned to be conceived of as the aggregate result of individual level decisions and events. These can be understood by focusing attention on the timing and sequencing of basic points in the life cycle: leaving school, leaving home, entering the workforce, getting married, becoming a household head, having children, moving to another neighborhood. All are inter-related and linked to the urban-industrial process through the family economy. Finally, the fourth area alters focus from processes that affect the entire population to the experience of particular groups. Here we examine how the urban-industrial experience affected important classes of people: blacks, Irish, Germans; the poor, welfare recipients and criminals; the new industrial elite and the aristocracy of labor; and women. Thus we learn how important human behavior — holding constant the urban-industrial setting — was mediated by the four major differentiators of experience: race, ethnicity, class and sex.

## The Significance of Data Manipulation in a Large Data Base

The utility of a large data base for interdisciplinary research comes not simply from the number of cases found in a given data set (though this is an advantage when, for instance, one examines four- or five-way tables), but from the number of data sets describing distinct characteristics of the larger whole being studied. It is not that there were thousands of individuals in Philadelphia and only hundreds in other settings, but that in addition to information describing these people there were data describing their families, neighborhoods, institutions, the local economy, transportation facilities, public and private services and the like. In other words, it is not the *length* of a given file, but the *breadth* of the entire research context that makes a data base sufficiently broad and detailed for interdisciplinary research. Such a data base enables us to move beyond static cross-sections and disconnected pieces of the urban experience that have necessarily characterized so much of previous urban historical research. The PSHP data base makes it possible to begin to see the *simultaneity*, the complex feedback loops, and the unanticipated consequences that issue from change in a given corner of the urban system. With the construction of the requisite research environment comes the ability to see — and thus raises the likelihood that we will understand — the interplay of personalities, political decisions, major events, institutional behavior and impersonal, socioeconomic and demographic forces. When research is approached in such a manner, it is likely that one's sense of causality will be radically altered as well, an inevitable consequence of seeing urban complexities more wholly.

In such a data-rich environment, data manipulation assumes a critical role. A computer-based research project may be divided into three distinct phases. The first is „Data Collection, Processing and Verification“. The last is „Data Analysis and Display“. „Data Manipulation“ is the intermediate stage generally involving the creation of new variables, the modification of file structure, and the aggregation of data within a file or the combination of data from different files to form new records. In a flow chart, data manipulation would stand between source files and statistical packages or mapping routines. Researchers who believe that they will be able to undertake analysis immediately after completing data verification — that is, after making certain that the conversion process has accurately transformed information from the historical source document to machine-readable form — are either in for a shock and considerable delay or they have quite limited data sets and narrow analytic concerns. When the research design is sophisticated and several large data sets are involved, data manipulation becomes a major challenge, often requiring the preparation of extensive new software specifically designed to perform a variety of data transformations.

My purpose in this section is not to elaborate fully on the data manipulation stage, but to identify two generic procedures within this stage that have the effect of making the whole greater than the sum of its parts. That is to say, that by combining the data sets in certain ways, researchers can undertake analyses not possible with the data in their original machine-readable form. The two procedures are record linkage and the summary of information by ecological unit.

### *Record Linkage*

The first of these is familiar to researchers engaged in studies of career and geographic mobility, and involves the identification of the same person in two or more separate files (without, of course, knowing in advance who these matched pairs will be). In the mobility studies, this procedure has almost always been an „over-time“ identification. The PSHP does record linkage over time not only of individual persons, but of individual families, business and manufacturing firms, and institutions. One of the sessions at the conference was devoted to this important topic.

However, I would like to explain the benefits that result from record linkage done for purposes other than longitudinal analysis: the identification and combination of records referring to the same persons, families, firms and organizational memberships across different data sets at the same point in time. While each data set was collected for specific analytic concerns that could be carried out without linking it to other data, when so combined they permit new and important analyses that could not be conducted working with each data set alone. Below are several brief examples of such data manipulation being undertaken at the PSHP.

(1) *City Street Directories to the Population Manuscript Schedules*: The population manuscript schedules did not include specific address until 1880. If the city street directories (which listed persons alphabetically by surname along with occu-

pation and street address) are machine-readable, then an automated record linkage program can be used to match persons found in both sources. In this operation, both name and occupation can be used to make positive identification without biasing results. Thus, drawing upon the socioeconomic and demographic information reported in the population census, the residential patterns of the city can be reconstructed and analyzed at the block or grid rather than at the ward level (with important benefits for subsequent longitudinal analysis because, unlike ward boundaries, the smaller areal units do not change over time).

(2) *Manufacturing Schedules to Business Directories*: The manufacturing manuscripts did not include street addresses for each firm and in 1880 they failed to report even ward number as had been the case in the earlier enumerations. Yet, with few exceptions, firms employing more than ten workers advertised their location in the city's business directories. Our automated record linkage program was used to match firms found in both sources. Thus it has been possible to undertake a series of detailed analyses of industrial geography and its significant relationships to a host of topics including residential patterns (which were dictated to a major extent by the locational decisions of manufacturing firms) and the „journey-to-work“.

(3) *The Journey-to-Work: Joining the Population and Manufacturing Schedules*: By overlaying the location of jobs (described in (2)) on the distribution of workers' residences (described in (1)), we were able to construct the labor shed parameters from which firms drew their employees and hence to derive the dimensions of the journey-to-work. In 1850 and 1880, respectively, 90 percent of the work force lived within four blocks (the median distance was two blocks) and within one-mile (the median distance was one-half mile) of the firms in which they worked. These results came from matching the industry-type coded for each firm in the manufacturing census with the same industry-type coded for each person's occupation in the population census. Using proximate spatial location and agreement in industry-type, it was possible in the absence of employee lists (only the number and sex of workers rather than actual names were reported in the manufacturing census) to associate specific people with the firms in which they were likely to have worked. The information describing industrial setting (e. g., the kind of firm — big or small, with or without mechanization, capital or labor intensive, etc.), can now be used, for example, as independent variables to explain the occupational mobility of individual workers.

(4) *Business Directories to Population Census*: When the business directories are linked to the population schedules, it becomes possible to study the socioeconomic and demographic characteristics of professionals, proprietors and craftsmen within the areas of the city in which they made their living. For instance, who were the owners of shops in the immigrant and black sections of city? More importantly, however, the study of the occupational universe, especially its stratification, is brought into far clearer focus. All studies of occupational structure and mobility — for that matter, the many studies that use occupation as a proxy for class or socioeconomic status — have been severely hampered by the limitations of the occupational data as reported in the population census. Significant numbers of ambiguous

job titles cannot be accurately assessed and coded. Does „liquor store“, for example, refer to the proprietor, the clerk or the clean-up man? Does „carpenter“, „baker“, or „tinsmith“ refer to the master craftsman or to the journeyman? When linked to the business directories, we are able to differentiate among these, to identify the proprietor and the master craftsman. We are thus in a position to classify the nineteenth-century occupational universe in terms far more relevant to it than is possible using the ahistorical stratification schema imported wholesale from contemporary sociological research.

(5) *Manufacturing Schedules to the Population Census*: Since incorporation required an act of the state legislature until 1875, the name of the manufacturing firm is almost invariably the name of its owner. When these people are traced to the population schedules it becomes possible to analyze variation in the industrial sphere in terms of the socioeconomic, demographic, and residential characteristics of the individual owners. For example, we can determine the origins of the new industrial elite: were they drawn from the ranks of the old commercial and native-born classes, from old handicraft manufacturing or from the ranks of the foreign-born and newcomers? Here is a quite distinct focus on the openness of the opportunity structure in nineteenth-century America. In the same manner, we can learn the background from which were drawn the owners of firms that produced capital or consumer goods, characterized different sectors of the economy, used mechanized production techniques, employed women and children, paid high or low wages, and so on down an even richer list when these categories are combined with each other.

(6) *Organizational Lists to the Population Schedules*: The PSHP is also making machine-readable extensive numbers of membership lists for a wide range of organizations: church-related, trade unions, political, self-help, leisure and literary. Work already completed for the black population (4 percent of the city's population), turned up 4,500 memberships, held by 3,000 persons, in 200 organizations. The stated purposes and prevailing types of organizations themselves provide important insights into the values, interests and needs of particular groups of people (differentiated by race, ethnicity, class and sex). When these lists are linked to the population census closest in time to the date of membership, however, it is possible to undertake a far reaching set of analyses. With new machine-readable files describing the socioeconomic, demographic and spatial characteristics of individual members, we can compare the profiles of members to non-members, of members and leaders across organizations, and of membership patterns over time. Thus we can infer a good deal about group values and attitudes; and when controlling for race, ethnicity and sex, we can explore the significance of socioeconomic class in determining internal group stratification.



## *Information Summarized by Ecological Unit*

The second generic type of data manipulation in a large and diversified data base requires the creation of a summary record for each ecological unit. The one piece of information common to all PSHP data is location in physical space, using as a coding category the roughly block square grid that serves as the project's basic areal unit. Such an effort involves not simply the combination of data from different files, but most frequently their transformation by aggregation to the level of the desired areal unit (a grid or combination of grids). Thus we can create a record that in some respects resembles (but is richer than) a contemporary census tract with an almost infinite number of variables drawn from the different PSHP source files.

The new records can be created in three basic ways: on an *ad hoc* basis; as a fixed set of predetermined variables; or by use of a generalized computer program that can be run and re-run. The first of these alternatives is the least desirable: it would require a considerable amount of computer programming each time a new areal summary file was created; it would make researchers who do not do their own computer programming always dependent on those who did; and it would be very expensive because it would be redundant. The second alternative would be relatively inexpensive, requiring only one long pass through all the files; but it would leave no room for change. This approach would be unable to accommodate the creation of new variables that emerge from ongoing research or that are required to answer new questions. The last alternative is the most attractive, though it, too, is initially very costly in software design and coding. Once available, however, it would enable any researcher to create files containing any desired variables, and it would be designed in such a way so that users would require no more programming knowledge than necessitated by the use of canned packages. The PSHP has chosen the last course of action, though the software is not yet completed.

The new files could then be used to support research into the relationship between individual behavior and the ecological setting by returning the historical actors to their relevant social context. Drawing on some of our research already conducted in intra-residential mobility, I would like to provide an example of how this approach works. Like most such studies, movement within the city was treated as a dichotomous dependent variable and the personal characteristics of movers and non-movers (e. g., age, occupation, birthplace, property holding, marital status) were used as independent variables. Using a multivariate approach that controls for the effect of highly intercorrelated independent variables, we were able to explain only ten percent of the observed variation in residential mobility. Such results pose serious problems in interpretation. Do the findings suggest that the decision to move is capricious? Or is it exceedingly complex, rooted in normative or psychological variables? Or is the explanation rooted in structural variables not included in the data set that describe the areas at both the origin and destination of the migration stream (e. g., the nature of the housing stock, the socioeconomic and demographic composition of the population, the availability of job opportunities, the presence of public services, institutions, and transportation facilities)?

When the new areal summary records are created, we will be in a position to answer these questions. First, both the new ecological files and those containing the individual „movers and stayers“ will be sorted by areal unit. Next a computer program will add to the machine-readable record of each person the ecological data for their origin and destination points. Finally, the new files will be re-analyzed using the multivariate routines, but this time there will be a host of new and potentially highly significant independent variables to supplement the explanatory process.

## Summary and Concluding Observations

The argument made in this essay is that collaborative interdisciplinary historical research is far easier to talk about than to operationalize successfully. Despite generations of praise and recommendations by countless scholars, active interdisciplinary research remains a goal, not an accomplishment. In the United States, structural factors — the system of hiring and rewards that govern the organization of research in institutions of higher learning — and cultural factors — that set of values that sanctify individual achievement while de-emphasizing collaborative effort — along with the non-applied uses of historical knowledge were offered as reasons why interdisciplinary efforts do not appear more frequently in American colleges and universities. A common data base of sufficient scope and detail was identified as a mechanism capable of supporting interdisciplinary research because it can facilitate the integration of effort by scholars from diverse social science backgrounds. Using the PSHP as a model for collaborative interdisciplinary historical research, we described the project's machine-readable data base, the kinds of analytic goals it supports, and several of the data manipulation strategies used to combine data sets thus enabling the attainment of analytic goals not possible when each information source was examined in isolation.

It is important to bear in mind that the shift in method and technique that characterizes the efforts of „new“ historians does not constitute a new form of research organization. Individual, disciplinary-enclosed research can continue now as it has in the past with change characterizing only the particular approach and tools, not the overall *organization* of research. In fact, this is by and large what happened in the U.S. Much of the historical research done could have been substantially improved if undertaken along with scholars from other disciplines, or at least in cooperation with colleagues in the same discipline. Because this was not done, individual researchers used different coding and categorization schema, different measures and variable values, and different record linkage techniques, all of which introduce serious bias and confound comparisons across studies. What is more, research remains fragmented; studies of family structure proceed independent of

those on the life course; studies of mobility are conducted without consideration of family behavior; both undertaken without regard for environmental setting or the impact of technological innovation. Quantitative data and methods will not by themselves, in other words, realize the opportunities made available by the technologies that facilitate much of our new research.

In closing I would like to offer the following suggestions to scholars beginning to undertake the new historical research. Do not embark on the process of data collection and processing as individuals. The constraints imposed on single researchers by the time and costs involved will severely limit the range of analytic possibilities, or if you will, the choice of dependent and independent variables in the research design. Instead, convene conferences and invite scholars from all disciplines who are interested in the new historical research. Place highest on the agenda the identification of the kinds of data necessary to support diverse analytic goals and the greatest number of researchers. Collect 100 percent samples wherever possible so that the many unanticipated uses of the data can be sustained by primary as well as secondary and tertiary researchers. Convert the historical information to machine-readable form without collapsing it into comprehensive categories; use basic codes that maintain a one-to-one relationship with the raw information so that the data can be recoded by subsequent researchers.

Pool your resources. Establish local, regional and national research centers. These centers need not at the outset be tied together by a tightly structured central theoretical framework; although some will argue that this must precede rather than follow the collection of the data, something far looser will suffice to get things moving. It is unlikely, moreover, that any meaningful agreement will be reached at this initial stage of research. Subgroups of researchers will emerge and more rigorous theoretical frameworks will evolve as the analytic work progresses. It appears that the establishment of the QUANTUM organization in Germany and the SSHA in the U.S. constitute important beginning steps in this direction and it augurs well for the future of interdisciplinary collaborative historical research.

## Appendix I

### *PSHP Research Associates (September, 1978)*

Name	Discipline and Affiliation	Research Interest
Burstein, Alan	Assistant Professor, Sociology, Washington University	Intra-urban Migration and Demographic Patterns
Cohen, Jeffrey	Ph. D. Candidate, History of Art, University of Pa.	Architectural History
Condran, Gretchen	Visiting Assistant Professor, Demography, University of Pa.	Nineteenth-Century Mortality
Cox, Harold	Professor, History, Wilkes College	Inter- and Intra-Urban Transportation Systems
Crum, John	Ph. D. Candidate, History, University of Delaware	Municipal Services and the Development of Bureaucracy
Ericksen, Eugene	Associate Professor, Sociology, Temple University	Industry, Residence and Com- munity Stability in 20th C. Phila- delphia
Feinman, Jay	Assistant Professor, Law, Rutgers University, Camden	19th Century Legal History
Fishbane, Richard	Ph. D. Candidate, Education, University of Pennsylvania	19th Century Urban Education History
Frisch, Michael	Associate Professor, History, Suny, Buffalo	Comparative Urban History
Furstenberg, Frank	Associate Professor, Sociology, University of Pa.	Family Structure and Behavior
Gillette, Howard	Assistant Professor, Dir. Amer. Studies, Geo. Washington Univ.	Nineteenth-Century Urban Political History
Glassberg, Eudice	Associate Professor, Social Wel- fare, Temple University	Benevolent Societies and the Response to Poverty
Goldin, Claudia	Assistant Professor, Economics, Princeton University	Family and Female Labor Force Participation

Name	Discipline and Affiliation	Research Interest
Greenberg, Stephanie	Sociologist, Ctr. for Study of Social Behavior, Research Triangle Institute	Industrial Location and Residential Patterns
Greenfield, Richard	Political Science, PSHP	Urban Form and Political Organization
Griffis, Jim	Ph. D. Candidate, Sociology, Temple University	Industrialization and Class Relationships
Gruenstein, John	Ph. D. Candidate, Economics, U of P; Federal Reserve Bank	Econometric Model for the Location of Manufacturing
Haines, Michael	Assistant Professor, Economics, Cornell University	Fertility, Mortality, Work and the Life-Cycle
Holley, John	Ph. D., Sociology and History, University of Edinburgh	Family Structure and Work in 19th C. England and America
Kawaguchi, Leslie	Ph. D. Candidate, History, U. C. L. A.	Formation of the German-American Community
Lane, Angela	Post Doctoral Research, Sociology, Temple University	Occupational Mobility and Industrial Structure
Laurie, Bruce	Assistant Professor, History, University of Massachusetts	Occupational Structure and Labor History
Lautzenhiser, Renee	Ph. D. Candidate, Geography, Pennsylvania State University	Geographical Distribution of Municipal Services
Light, Dale	Ph. D. Candidate, History, University of Pennsylvania	The Formation of the Irish Community
Miller, Roger	Lecturer, Reg. Science and Urb. Studies, U. of Pa.	Application of „Time-Geography“ to Nineteenth-Century Cities
Modell, John	Professor, History, University of Minnesota	Life Course Development and Urban Growth
Potman, Stephen	Assoc. Professor, City and Regional Planning, U. of Pa.	Work, Residence and Transportation Modelling
Roberts, Jeffrey	Ph. D. Candidate, History Temple University	Development of the Central District and Hist'l Geography

Name	Discipline and Affiliation	Research Interest
Rosenberg, Charles	Professor, History, University of Pennsylvania	19th Century Medical and Social History
Schmitz, Mark	Assistant Professor, Economics, University of Delaware	Manufacturing Productivity
Seaman, Jeffrey	Sociologist, PSHP and SPUP	Methodology, Statistics, Demography
Siry, Joseph	Ph. D. Candidate, Architecture, University of Pennsylvania	Suburban Housing Development and Architectural History
Steinberg, Allen	Ph. D. Candidate, History, Columbia University	Poverty and Criminality
Stolley, Paul	Professor, Research Medicine, University of Pennsylvania	Epidemiology and Mortality
Thomas, George	Architectural Historian	Architectural History
Ulle, Robert	Ph. D. Candidate, History, University of Pennsylvania	Culture and Institutions Among Black Philadelphians
Wallock, Leonard	Ph. D. Candidate, History, Columbia University	Artisans and Industrialization
Whitney, William	Lecturer, Economics, University of Pennsylvania	Location and Productivity of Manufacturing
Yancey, William	Associate Professor, Sociology, Temple University	Industry, Residence and Community Stability in 20th C. Philadelphia

## Appendix II

### *PSHP Papers published and unpublished*

Burstein, Alan N.,

Intra-Urban Migration Patterns of Irish and German Immigrants, paper presented at the Social Science History Association (SSHA) Convention, Philadelphia October 1976; to appear as *Immigrants and Residential Mobility: The Irish and Germans in Philadelphia, 1850 to 1880\**.

Patterns of Segregation and the Residential Experience, in: *Historical Methods Newsletter*, 9, No. 2-3 (March-June)\*.

Drobis, Susan M.,

Occupation and Residential Differentiation: A Historical Application of Cluster Analysis, in: *Historical Methods Newsletter*, 9, No. 2-3 (March-June 1976).

Eisenhardt, Miriam, Sultanik, Jeffrey, and Berman, Alan,  
The Five Irish Clusters in 1880 Philadelphia, May 1974.

Fishbane, Richard B.,

'The Shallow Boast of Cheapness': Public School Teaching as a Profession in Philadelphia, 1865-1890, in: *Pennsylvania Magazine of History and Biography*, January 1979.

Furstenberg, Frank, Hershberg, Theodore, and Modell, John,

The Origins of the Female-Headed Black Family: The Impact of the Urban Environment, in: *Journal of Interdisciplinary History*, 6, No. 2 (September 1975). Reprinted in Staples, Robert (ed.), *The Black Family: Essays and Studies*, 2nd ed., Belmont/Calif. 1978\*.

Furstenberg, Frank, Hershberg, Theodore, Modell, John, and Strong, Douglas,

What Happened When the Census was Redone: An Analysis of the Recount of 1870 in Philadelphia, paper presented at the Brigham Young University Family Research Conference, Historical Change in Marriage and the Family, Provo/Utah, March 1978.

Gillette, Howard,

The Emergence of the Modern Metropolis: Philadelphia in the Age of Its Consolidation, in: Gillette, Howard, and Cutler, William (eds.), *The Divided Metropolis: Social and Spatial Dimensions of Philadelphia 1820-1940*, forthcoming 1979.

Glassberg, Eudice,

Work, Wages and the Cost of Living: Ethnic Differences and the Poverty Line, Philadelphia, 1880, in: *Pennsylvania History*, January 1979.

Goldin, Claudia,

Household and Market Production Functions in a Late-Nineteenth Century American City: Philadelphia, 1880, in: *Explorations in Economic History* (forthcoming).

- Goldin, Claudia,  
Family Strategies in Late-Nineteenth-Century Philadelphia\*.
- Greenberg, Stephanie W.,  
Industrial Location and Ethnic Residential Patterns in an Industrializing City, Philadelphia, 1880, paper presented at the 1976 meeting of the SSHA, Philadelphia/Pa.\*.
- The Relationship between Work and Residence in an Industrializing City: Philadelphia, 1880, in: Gillette, The Divided Metropolis.
- Industrial Location and the Residential Patterns of Occupational Groups: Philadelphia, 1880, SSHA (November 1978).
- Greenberg, Stephanie W., Hershberg, Theodore, Whitney, William, and Roberts, Jeffrey,  
The Determinants and Dynamics of Intra-Urban Manufacturing Location: A Perspective on Nineteenth-Century Philadelphia, SSHA (November 1978)\*.
- Gruenstein, John,  
Components of Manufacturing Location Change, 1850–1880, SSHA (1976).
- Haines, Michael,  
Poverty, Economic Stress and the Family in a Nineteenth-Century American City: Philadelphia, 1880\*.
- Hershberg, Theodore,  
Free Blacks in Antebellum Philadelphia: A Study of Ex-Slaves, Free-Born and Socioeconomic Decline, in: Journal of Social History, 5, No. 2 (December 1971), pp. 183–209. Reprinted in Haller, Mark, and Davis, Alan (eds.), The Peoples of Philadelphia, 1973; Katz, Michael B. (ed.), Education and American Social History, New York 1973; Genovese, Eugene, and Miller, Elinor (eds.), Plantation, Town and County, Urbana 1974; and the Bobbs-Merrill Reprint Series in American History, Indianapolis 1973\*.
- Free-Born and Slave-Born Blacks in Antebellum Philadelphia, in: Genovese, Eugene, and Engerman, Stanley (eds.), Slavery and Race in the Western Hemisphere, Princeton 1975.
- Towards the Historical Study of Ethnicity, in: Journal of Ethnic Studies, 1, No. 3 (Spring 1973).
- A Method for the Computerized Study of Family and Household Structure Using the Manuscript Schedules of the U.S. Census of Population, 1850–1880, in: The Family in Historical Perspective, 1, No. 3 (Spring 1973).
- Hershberg, Theodore, and Williams, Henry,  
Mulattoes and Blacks: Intra-Group Color Difference and Social Stratification in Nineteenth-Century Philadelphia (1974)\*.
- Hershberg, Theodore,  
The Organization of Historical Research, in: AHA Newsletter, 12, No. 7 (October 1976).
- Hershberg, Theodore, Cox, Harold, Light, Dale, and Greenfield, Richard,  
The „Journey-to Work”: An Empirical Investigation of Work, Residence and Transportation,



Philadelphia, 1850 and 1880 (1974)\*.

Hershberg, Theodore, Katz, Michael, Glasco, Laurence, Blumin, Stuart, and Griffen, Clyde,  
Occupation and Ethnicity in Five Nineteenth-Century Cities: A Collaborative Inquiry, in:  
Historical Methods Newsletter, 7, No. 3 (June 1974)\*.

Hershberg, Theodore,  
The Philadelphia Social History Project: An Overview and Progress Report, in: Review of  
Public Data Use, 4, No. 1 (January 1976).

Hershberg, Theodore (ed.),  
Toward an Interdisciplinary History of the City: Work, Space, Family and Group Experience  
in Nineteenth-Century Philadelphia, forthcoming 1979. All essays identified with an „\*\*“  
will appear in this volume.

Hershberg, Theodore,  
The Philadelphia Social History Project: An Introduction, in: Historical Methods Newsletter,  
9, No. 2-3 (March-June 1976).

Hershberg, Theodore, and Dockhorn, Robert,  
Occupational Classification, in: Historical Methods Newsletter, 9, No. 2-3 (March-June  
1976).

Hershberg, Theodore,  
The Historical Study of Urban Space, in: Historical Methods Newsletter, 9, No. 2-3 (March  
-June 1976).

Hershberg, Theodore, Burstein, Alan, and Dockhorn, Robert,  
Record Linkage, in: Historical Methods Newsletter, 9, No. 2-3 (March-June 1976).

Hershberg, Theodore,  
The Urban History: An Interdisciplinary History of the City, prepared for a PSHP-sponsored  
conference (Philadelphia: August 1-4, 1977); published as The New Urban History: An In-  
terdisciplinary History of the city, in: Journal of Urban History, 5, No. 2 (November 1978).

Hershberg, Theodore, Ericksen, Eugene, Yancey, William, Greenberg, Stephanie, and Burstein,  
Alan,  
A Tale of Three Cities: Blacks and Immigrants in Philadelphia, 1850-1880, 1930 and 1970,  
for a special issue on Race and Residence in American Cities, in: The Annals, 440 (January  
1979).

Laurie, Bruce,  
Radicals, Revivalists, and Traditionalists: Class and Culture in Antebellum Philadelphia (MS.).

Laurie, Bruce, Hershberg, Theodore, Alter, George,  
Immigrants and Industry: The Philadelphia Experience, 1850-1880, in: Journal of Social  
History, 9, No. 2 (December 1975); this essay, presented to the Sixth International Congress  
on Economic History (Copenhagen, 1974), was reprinted in Richard Ehrlich (ed.), Immi-  
grants in Industrial America: 1850-1920, Charlottesville/Va. 1977\*.

- Laurie, Bruce, and Schmitz, Mark,  
 Manufacture and Productivity: The Making of an Industrial Base\*.
- Lees, Lynn, and Modell, John,  
 The Irish Countryman Urbanized: A Comparative Perspective on the Famine Migration, London and Philadelphia, in: *Journal of Urban History*, 3, No. 4 (August 1977)\*.
- Miller, Roger,  
 Household Activity Patterns in the Emerging Nineteenth-Century Suburb, *SSHA* (November 1978).
- Miller, Roger, and Siry, Joseph,  
 The Emerging Suburb: West Philadelphia, 1850–1880 (1978).
- Modell, John, Furstenberg, Frank F., and Hershberg, Theodore,  
 Social Change and Transitions to Adulthood in Historical Perspective, in: *Journal of Family History*, 1, No. 1 (Autumn 1976). Reprinted in Gordon, Michael (ed.), *The American Family in Social-Historical Perspective*, 2nd ed., New York 1978, pp. 192–219.
- Modell, John, Furstenberg, Frank, and Strong, Douglas,  
 The Timing of Marriage in the Transition to Adulthood: The Continuity of Change, 1860–1975, in: *American Journal of Sociology*, special issue on the history of family life, 1978.
- Modell, John,  
 Suburbanization and Change in the American Family: Their Connection, in: *Journal of Interdisciplinary History*, forthcoming 1979.
- Putman, Stephan H.,  
 Calibrating a Disaggregated Residential Allocation Model-DRAM, in: *London Papers in Regional Science* (1977).  
 Calibrating a Residential Model for Nineteenth-Century Philadelphia, in: *Environment and Planning, A*, 9 (1977).
- Roberts, Jeffrey,  
 The Central District, *SSHA* (1976).  
 Railroads and the Growth of the Downtown: Philadelphia, 1830–1900, in: Gillette, The Divided Metropolis.
- Sobotowski, Christopher,  
 Software, in: *Historical Methods Newsletter*, 9, No. 2–3 (March–June 1976).
- Thomas, George,  
 Architectural Patronage and the Shape of the City, 1850–1925, in: Gillette, The Divided Metropolis.
- Whitney, William,  
 Manufacturing Location at Mid-Century, *SSHA* (1976).

Williams, Henry, and Uile, Robert,

Frankford, Philadelphia: The Development of a Nineteenth-Century Urban Black Community, in: *Pennsylvania Heritage* (December 1977).

### *Ph. D. Dissertations*

#### Completed:

Hershberg, Theodore,

The Philadelphia Social History Project: A Methodological History, Ph. D. Dissertation, History, Stanford University 1973.

Burstein, Alan N.,

Residential Distribution and Mobility of Irish and German Immigrants in Philadelphia, 1850–1880, Ph. D. Dissertation, Demography, University of Pennsylvania 1975.

Greenberg, Stephanie,

Industrialization in Philadelphia: The Relationship Between Industrial Location and Residential Patterns, 1880–1930, Ph. D. Dissertation, Sociology, Temple University 1977.

Miller, Roger,

Time-Geographic Assessment of the Impact of Horse-Care Transportation in Philadelphia, 1850–1860, Ph. D. Dissertation, Geography, University of California, Berkeley, 1979.

#### In Process:

Cohen, Jeffrey,

Housing Philadelphians: The Relationship between Changing Architectural Types and Social and Spatial Differentiation, Architectural History, University of Pennsylvania.

Crum, John,

Development of Municipal Services in Nineteenth-Century Philadelphia, History, University of Delaware.

Fishbane, Richard,

Teachers in Transition: The Development of Public School Teaching as a Profession in Nineteenth-Century Philadelphia, Social Foundations of Education, University of Pennsylvania.

Glassberg, Eudice,

Philadelphians in Need: Client Experiences with Two Philadelphia Benevolent Societies, 1830–1880, Social Work, University of Pennsylvania.

- Griffis, James,  
The Development and Reproduction of Class Relations: Philadelphia, 1850–1930, Sociology, Temple University.
- Gruenstein, John,  
Dynamic Models of Employment and Population Location for Philadelphia, 1850–1880, Economics, University of Pennsylvania.
- Kawaguchi, Leslie Ann,  
The Germans in Philadelphia During the Nineteenth-Century, History, University of California, Los Angeles.
- Lautzenhiser, Renee,  
Development of Municipal Services: Water, Sewage, Gas, Fire and Police: Philadelphia, 1850 to 1880, Geography, Pennsylvania State University.
- Light, Dale,  
The Evolution of the Irish Community in Philadelphia, 1850–1880, History, University of Pennsylvania.
- Roberts, Jeffrey,  
The Growth of Philadelphia's Central District, 1830–1900, History, Temple University.
- Steinberg, Allen,  
Crime, Criminals, and the Meaning of the 'Dangerous Class', History, Columbia University.
- Ulle, Robert,  
Institutional Development and Attitudes in Nineteenth-Century Black Philadelphia, History, University of Pennsylvania.
- Wallock, Leonard,  
The Impact of Industrial Capitalism on the Skilled Sector of the Workforce: Philadelphia's Artisans in the Nineteenth Century, History, Columbia University.
- M. S. Theses: Graduate Group in Computer and Information Sciences, Moore School of Electrical Engineering, University of Pennsylvania:
- Miller, Buffington Clay,  
A Computerized Method of Determining Family Structure from Mid-Nineteenth Century Census Data (August 1972),
- Palladinetti, Stephen J.,  
The Evolution of the Source Conversion Process at the Philadelphia Social History Project (May 1974).
- Stickel, Terry,  
A Computerized Nominal Record Linkage System (December 1972).

## Appendix III

### *Machine-Readable Data Describing Philadelphia County*

#### I. Areal Unit

A grid pattern, with each unit one by one-and-one quarter blocks (660' X 775'), was imposed on the map of the entire county (130 square miles) producing ca. 7,100 total grids.

All the machine-readable information identified below has been (or will be) coded for precise grid location; therefore it is possible to create an extensive set of variables for use in ecological analyses.

#### II. U.S. Population Manuscript Census Schedules: 1850, 1860, 1870, 1880

For the study of individuals and groups: Blacks: all persons (94,000); Irish: all males 18+ (142,000); Germans: all males 18+ (84,000); native-white-Americans: sample males 18+ (50,000). Longitudinal files (between-census linkages) of individual Black males and females, Irish males, and German males have been created.

For the study of family structure: 2,000 households, with ca. 6 persons per household for each ethnic group (Irish, German, native-white-American) in each census year. Total sample 144,000.

Variables for each of 500,000 individuals (differ slightly by census year): name, address, age, sex, race, color, occupation, real and personal property, place of birth, literacy, material status, school attendance.

For the study of neighborhood: Total population: 1850—408,000; 1860—565,000 1870—647,000; 1880—840,000. For each inhabitant of the 7,100 grid units, we record the age (8 categories), sex, and ethnicity (9 categories including 2nd generation in 1880).

#### III. Pennsylvania Abolition Society and Society of Friends Manuscript Census Schedules: 1838, 1847, 1856.

Data describes 11,500 households — all black households in Philadelphia County.

Variables for each household head and his household include (differ slightly by census year): name, sex, status-at-birth, occupation, wages, real and personal property, literacy, education, religion, membership in beneficial societies and temperance societies, taxes, rents, dwelling size.

#### IV. U.S. Manufacturing Manuscripts Census Schedules: 1850, 1860, 1870, 1880

Data describes ca. 29,000 individual firms (100% sample). Firms: 4,700 (1850); 8,500 (1880).

Variables: name of firm; type of product; amount of capital investment; type of power; type and number of machines; number of employees by sex; average wages paid by sex; number of months in operation each year; raw materials: kinds, quantities, values; finished products: kinds, quantities, values; and address and grid location.

- V. **City Business Directories: 1850, 1860, 1870, 1880**  
 Data describes ca. 127,000 individual firms (100% sample). Firms: 10,000 (1850); 28,000 (1860); 39,000 (1870); 47,000 (1880).  
 Variables: name of subscriber; type of product or service; and address and grid location.
- VI. **Transportation Network: 1840–1880**  
 Data describes a total of 150 routes including all forms of transportation: street, railways, railroads, trolleys, omnibuses.  
 Variables: date of incorporation; company name; date of merger; reconstruction of block-by-block routes and conversion of precise „X“–„Y“ coordinates to machine-readable form.
- VII. **Mortality Register: July 1, 1880–June 30, 1881**  
 Data describes ca. 19,000 individuals who died in Philadelphia during the twelve-month period following the Bureau of the Census (June, 1880) enumeration of the city's population .  
 Variables: name, color, sex, age, material status, date of death, cause of death, attending physician, occupation, birthplace, parent's names (if a minor), ward, address, place of burial.
- VIII. **Sewage Facilities: 1867–1885**  
 Data includes the following variables: year of construction of sewer line; street; ward; dates of authorization, approval and final estimate; length; assessment; cost to city and property owners; builder; grid location.
- IX. **Institutions (public, private, secular, religious): 1850, 1860, 1870, 1880**  
 Data describes 3,000 institutions.  
 Variables: name, year, type, address, grid location.
- X. **Voluntary Associations: 1820–1900**  
 Data includes Black, Irish, and German individuals who were members of various Philadelphia organizations. Blacks: 3,200 individuals, 4,500 memberships, 200 organizations. Irish: 3,500 individuals, 6,000 memberships, 350 organizations. Germans: 2,500 individuals, 5,000 memberships, 200 organizations.  
 Variables: name, ethnicity, organization, position, year of membership.  
 In addition to this organizational data, our files include census information for those people we have linked to the four U.S. and three Quaker and Abolitionist censuses. To date, we have completed linkages for the Black associational members. Extensive lists for the Irish and German are in preparation; individual members will be linked to four U.S. Censuses.

Conditions and Chances in the Computer-Aided Analysis  
of Historical Record Series: The Case of Early Modern  
Urban Administration

This report has been written for a twofold purpose. First, it provides some information on a research project of the Sonderforschungsbereich 8, University of Tübingen, the „cities-project“, dealing with social inequality in selected towns of South Germany and with the prosopographical characteristics of their political elites<sup>1</sup>. Second, the report will present some ideas concerning the reinterpretation of the output of early modern urban administration which point to new questions and research results in early modern urban social history. It focuses on the methodological conditions and opportunities of the cities-project carried out under the official title „Towns in Late Medieval Times and During the Reformation Period in South Germany“ and employing quantitative methods using EDP and certain approaches of the social sciences.

Subjects of investigation are ten early modern towns which have been selected according to typological criteria such as size, legal status, territorial affiliation („Stadtherrschaft“), function, attitude toward the Reformation movement, and of course, the sources available. This project is part of a larger research program evaluating the impact of non-theological factors on the Reformation in Southern German cities, imperial or territorial. In addition to this subsidiary function, the „cities-project“ has its own legitimation in the effort to analyse a larger number of early modern towns on the basis of one methodological approach and in comparative perspective.

The end of the project would not be exciting without the given temporal frame. The study is focused upon towns of the 15th and 16th centuries, and it is just this fact which defines the specific conditions, difficulties, problems, and opportunities of the research done.

Until today, historical research was concentrated, as far as it used complex statistical methods and EDP, on the social and economic history of the 19th and 20th centuries. This branch of historical work which has been carried out in the German Federal Republic in the last few years is known as historical social science („Historische Sozialwissenschaft“). In regard to early modern social history the situation is

<sup>1</sup> Ingrid Bátori, Dieter Demandt, and the author are participating in the „cities-project“ under the leadership of Hans-Christoph Rublack.

quite different. As Moeller, Dickens, and Ozment pointed out, there is little doubt of the very important role the cities played in the Reformation; for example, Dickens called the Reformation „an urban event“<sup>2</sup>. Beyond this, many studies, in fact case studies on individual towns, have been published in recent years. However, early modern urban social history shows at least two deficits, one methodological and the other theoretical.

- Traditional idiographic methods have dominated early modern social history and they do so today emphasising the individuality and singularity of each case.
- Early modern urban social history did not develop characteristics and criteria fitting through systematic comparative analysis, or did so in an insufficient way. There was and is little interest in the application, testing, and if needed, modification of theoretical models and/or middle range theories as they are worked out, discussed and used by the social sciences<sup>3</sup>.
- Early modern urban social history did not see the possibilities and advantages of of quantitative analysis or did so on an unsatisfactory level. Until today, the general standard of statistical data processing was extremely low.
- There is still a remarkable unwillingness to use technical means like the computer in early modern urban social history.

In short, the social history of early modern cities, as Leo Schnore put it, remains „a field without focus“. It runs the risk of not meeting the more developed standards of the social sciences and historical social science for the post-industrial eras. The „cities-project“ of the SFB 8 tries to compensate for some of these deficits to a certain and limited extent, at least within the possibilities of its thematic frame, and according to the available sources and the professional training of the collaborators who are involved.

As a starting point we assume that early modern towns with their relatively permanent legal, social, economic, and social-psychological relationships represent a complex social unit which can be called a social system. The structure of these social systems can be subdivided, described and analysed in accordance with certain criteria and dimensions. As Schnore has pointed out, the most important facets of the urban social systems are the demographic, ecological, structural and socialpsychological aspects<sup>4</sup>. In addition, a systematic investigation is bound to look at its results both in a macro-analytical and a micro-analytical perspective.

As a general principle it is impossible to explore „past present times“ (N. Luhmann) unless they have produced and handed down adequate and sufficient sources. This is the case on the whole, as far as the selected towns are concerned. (In visiting more than 120 archives in Southern Germany we found in about 20 to 25 ar-

<sup>2</sup> Dickens, G. A., *Luther and The German Nation*, London 1974, p. 182.

<sup>3</sup> See now as an extremely valuable exception Brady, Thomas A., *Ruling Class, Regime and Reformation at Strasbourg 1520–1555*, Leiden 1978 (= *Studies in Medieval and Reformation Thought*, XXII).

<sup>4</sup> Schnore, L. F., *Problems in the Quantitative Study of Urban History*, in: *DYOS*, (1971), pp. 189–208.



chives material and record series which provide a worthwhile basis for research on topics including the policy of urban authorities towards ecclesiastical institutions and a comparative study of the patterns of the Reformation-process.)

The critique of historical sources subdivides the corpus of sources into tradition („Tradition“) and record series („Überreste“); the latter being of great importance to the present research project in discussion. Of the record series in view, one deals primarily with the output of urban administrations and bureaucracies which German historical science generally calls „Akten“ (records). Records in this sense are characterized not by the outer form, in contrast, for example, to the medieval charters and documents, but by their function and their contents. The late medieval, early modern records have the function of being a „Memorialschriftstück“ (Pitz), i. e. documents that testify to the execution of administrative acts prepared for the use of the acting authority. As a consequence, investigation and interpretation of record series must start with acquisition of knowledge of the respective early modern bureaucratic organizations and their activities. In contrast to modern standards, early modern record series are not interpretable easily or readily, they demand intimate knowledge of the structure and scope of duties and functions of the administrative authorities in general and of municipal administrations specifically. In this context administrative activities mean, once again in contrast to modern times, all activities of the municipal authorities, whether they belong to the administrative field in a strict sense, or to the field of legislative authority or the judicial area in modern terms<sup>5</sup>.

Now, late medieval and early administrative organization presents an extremely complicated and confused picture of entirely unsettled administrative bodies and structures corresponding in no way to the modern administrative and bureaucratic system. In addition, local and regional peculiarities complicate these general facts considerably. The auxiliary science dealing with historical records has detected the basic principle guiding the late medieval and early modern urban constitution and administration: they tend to a perfect unity and centralization of the urban governmental power in the hands of the municipal council with simultaneous delegation of certain administrative functions to special administrative bodies. The constitution of the municipal council defines the field of urban administrative activities and, vice versa, the administrative activities specify the municipal constitution. The council of the towns in the 15th and 16th centuries differs from its medieval forerunners as from modern urban administration by its fundamentally unlimited administrative power. Within the urban administration, the council as a whole governed and acted in a nearly arbitrary way; it was not at all limited by rights of administrative or bureaucratic rules. In those days, municipal administration was a field without legal basis. Individual clerks executed certain administrative functions as a duty, not as right. The all embracing responsibility and power of the municipal council has some

<sup>5</sup> Pitz, E., *Schrift- und Aktenwesen der städtischen Verwaltung im Spätmittelalter*, Köln-Nürnberg-Lübeck, Beitrag zur vergleichenden Städteforschung und zur spätmittelalterlichen Aktenkunde, Köln 1959, p. 29.

characteristic consequences: as a general principle no restriction of administrative activities and functions; no defined areas of powers; no sequence of courts; no or only minimal disputes about competences. The general and everyday conditions and needs of life in urban social systems, as I would like to emphasize: exclusively these conditions and needs, induced and necessarily modified style, procedures, range, and topics of municipal administration. In spite of the tendency to unity and centralization, the internal development of the social systems and the emergence of new external political conditions caused a growing differentiation of the fields of administrative activities and functions within the towns of the 13th and 14th centuries at the latest. They necessitated an effective execution and handling of specific financial and military, economic and commercial acts of administration by the municipal authorities; since the beginning of the Reformation movement additional administrative functions and needs grew up in the domain of social security, church, and school. But this development did not break the general principle of the late medieval and early modern administrative structure: city council and the existing specialized administrative boards have to be regarded as *one* bureaucratic organization, in spite of all varieties in local municipal administration, whereas today administration represents a more or less clear hierarchy of complex departments and purviews with early important autonomous free play and powers.

In a historical perspective, the necessity to administrate is older than the practice of doing so in writing. To be sure, for the purpose of our research the general realization of writing administration since the 14th and 15th centuries indicates the point of take off for a systematic research on early modern administration and its output. With that, the most relevant universal conditions for the analysis of historical record series by the early modern urban social history are fixed: They depend on the knowledge of the nature of the early modern administrative and municipal organization in general and the local and regional varieties in particular; they depend on the knowledge of the historical situation given, i. e. the researcher has to pay adequate attention to the dimensions of time and space; they depend on a sufficient training in auxiliary scientific techniques and methods such as source-critique, paleography, early modern high or low German language, humanistic latin, history of money, numismatics, history of weights and measures, and so on. Last not least they depend on the existence of usable record series in sufficient density and quality.

In the case these requirements are fulfilled, historical record series may be interpreted as process-produced historical data. They have the following characteristics:

- Process-produced historical data have been produced without the intention of historical information, but for certain official, commercial or private ends or without any observable purpose.
- Above all, they are results of administrative and/or bureaucratic acts in a broader sense.
- They are topical, that means up-to-date in the sense of passed presence.
- They contain large quantities of mostly homogeneous information.
- As a rule, they are quantitative or can be quantified.

- They inform directly on facts of passed presences which I would like to call „function of direct information“ and
- they give evidence in an indirect way as indicators on topics whose notification was neither the aim nor the purpose of the data-producer; this I would like to call „function of indirect information“. (I am not sure whether it is useful to make the distinction of „hard“ respectively „soft“ data.)
- They can be processed, i. e., they are feasible for the systematic and automatic analysis of properties and connections, especially by compromising, manipulation and reduction of their manifest or latent information.
- They are to be subjected to an adequate historical critique before being processed and interpreted.

The „cities-project“ of the SFB 8 concentrates, besides other towns, on the town of Kitzingen on the Main river. This town may serve the purpose of demonstrating the concrete realization of the research work going on.

In the 15th and 16th centuries the town of Kitzingen was pledged to the county (Markgrafschaft) of Brandenburg-Ansbach. It was not an autonomous imperial town, but a so called „landsässige“ town under a secular lord. Compared with the number of big or bigger cities of the Holy Roman Empire, Kitzingen was a relatively unimportant example of middle-sized town with about 3,000 inhabitants earning their living mainly by viniculture. The level of complexity of this urban social system does not match the niveau of important urban centers as Strasbourg, Nürnberg, Augsburg, Cologne or Lübeck. But in our context the town of Kitzingen is a special case in so far as it is outstanding owing to an excellent situation of its record series. Despite the simple and rough structure of the Kitzingen administration, the abundance and the quality of the record series handed down reflect an extensive and precise administrative activity of the municipal council. I will describe the most important record series and then give an outline of their processing and the substantial results coming forward.

The development of the municipal constitution went along with the emergence and formation of an efficient administration of finances. The salient characteristics of the administration of the finances — such as autonomy, regular direct taxation of all citizens respectively productive system of indirect taxes, taxation as a forced impost, principle of nonaffectation, audit by the council and/or the citizenship and so on — invited the comparison with modern administration of taxes. The records produced by this urban administration of finances contain information which are process-produced historical data in a typical way. As a rule the date can be found in tax-roles or registers of taxation; they were used as lists of debit and credit with the purpose of fixing and collecting taxes. The tax roles of the town of Kitzingen do not only give information about the taxation of the citizen (as far as they were obliged or worthy to pay taxes) and the tax yield of the town as a whole, moreover they register the whole taxable property of all persons. Therefore they enable the researcher to investigate in a penetrating way the socio-economic situation of the town and the citizenship. It is possible to compute, within certain tolerances, the number of inhabitants. In the case of the Kitzingen tax-roles which are subdivided

in quarters, it is possible to give exact information about the number, the density, the displacement of population and properties in every quarter throughout the century. Unfortunately, the registration of the occupations seems not to be complete, although it is not impossible to reconstruct approximately the occupational structure of the town. Finally, tax-roles record the number of unmarried women and the widows.

The parish registers of the town of Kitzingen give the best information about the demographic facts. They have a complicated and rather confused coherence which is just a reflection of the unarranged and not yet fully formed administrative structure of the early modern town<sup>6</sup>. The historical auxiliary science did detect the inner dependence and tangle of the parish registers, so, in principle, there are enough data to deal with the demographic aspect of this town. For the time being, the „cities-project“ confines itself to analyse the demographic data for the political elite. Complete analysis of all data on birth, marriage and death will follow at a later date.

In connection with the end of the pawning of the town in 1628/29 two stout volumes were produced containing data which have an inestimable value for the analysis of the political elite, the members of the council, the structure of the municipal administration and so on. The so called compilation of Paul Rücklein, a town-clerk, and the *Topographia* of Salomon Codoman, a rector, contain nearly complete lists of all municipal officials and functionaries from the beginning of the 15th century up to 1628. Again, the correct interpretation requires a precise knowledge of the urban constitution and the administrative apparatus which on the other side depends on the accurate analysis of the data to be found in these volumes. This dialectic method of interpreting and evaluating data by the acquaintance with the administrative activities which produced them and which can mainly be studied by just investigating those data can be regarded as a typical example of the scientific intercourse with process-produced historical data. Other historical sources like chronicles are helpful and — au fond — indispensable means to fulfill the aims of historical research.

It is neither possible nor necessary to present here all the other record series completely. The catalogue of records we have been dealing with includes such important series as the books and registers containing the marriage contracts or contracts of guardianship (about 6,000 pages), the minute-books of the municipal council, beginning in 1527 or the so called „Steuer Receßbuch“, a sort of account-book enabling the reconstruction of the urban budget for over 30 years. A rough computation of the size of the existing record series brings to light that in Kitzingen there are about 12 000 to 14 000 pages with about 100 000 data relevant for the

<sup>6</sup> See Simon, M., *Die Kirchenbücher Kitzingens und ihre Stellung in der Geschichte der Kirchenbuchführung*, in: *Zeitschrift für bayerische Kirchengeschichte*, 26 (1957), pp. 146–162. The parish registers are to be found in the Archives of the Katholisches Pfarramt St. Johannes and the Evangelisches Dekanat, both in Kitzingen. All other mentioned sources are in the Stadtarchiv Kitzingen.

cities-project. So, even this case of a small insignificant early modern town confirms David Herlihy's statement that it is possible (I would like to add: necessary) to use quantitative methods including EDP at this early stage of the „protostatistical era“ with great success.

For the purpose of a systematic analysis some of the mentioned record series were brought into machine readable form, completely or partly, and stored on tape. A number of specific computer-programs, written in FORTRAN, manipulate the data in order to print alphabetical lists, lists of taxation ranks and graphs of the distribution of frequencies. SPSS is used for statistical computations (measures of localisation, dispersion, skewness, curtosis etc). Only in a few cases it seemed pithy and suggestive to compute more complex statistics like measures of regression and correlation. The efforts to aggregate data relevant and needed for the prosopography of the political elite automatically, i. e. to manage record linkage corresponding to the needs of the „cities-project“ did not work as requested; but as a matter of fact, the work, done by nonprofessionals, is not yet finished.

The systematic analysis of all record series available, even if computer-aided, does not yield a perfect analysis of the social structure of the town of Kitzingen, at least if the standards of the social sciences are requested. But it is possible to describe and analyse some facets of the urban social system covering mainly, in Leo Schnore's terms, the demographic, ecological, and structural aspects of the urban history. The contrasting of social history and biographical history is the focus of the research done; that means methodologically, the confrontation of data and issues obtained by macro-analytical approaches and techniques on the one hand and data and results obtained in a micro-analytical way on the other hand. This corresponds to the basic hypothesis that urban leading classes do necessarily function within a social context. Social and political leadership is conceivable only in a process of social interdependence. There are no leaders without society, no elite without masses in the early modern urban social systems, too, although leadership in the 16th century within an urban context is different from leadership in a post-industrial mass society. Consequently, each investigation of political and social leaders has to throw light on the range and the structure of the urban social system which is as a whole the basis to define the leading class. Because of the direct and narrow coherence of elite and social differentiation the analysis of social inequalities is of great importance. The „cities-project“ interprets social differentiation as socio-economic stratification on a vertical continuum. Social appreciation and prestige are a necessary, but not exclusive criteria to define the term. The gradation of the social strata is possible according to the contemporary ideas of rank and status of „Stände“ and other social groups given. But there is no doubt that the „Zeitgeist“, emerging from those ideas, was the spirit of the masters and the lords or, as Engels and Marx said, the ruling ideas of a time were always and only the ideas of the ruling classes. Conceptions and clues of common equality did exist, but in the contemporary corset of a given unequal world which could not be called fundamentally in question. Common equality and common best („gemeiner Nutzen“) were a privilege of the companions respectively of the citizens owning the full municipal citizenship, and

they were always and only a part of the social system, although they have been usually identified with the social „organism“ as a whole by a traditional historiography. So, we have to ask, whether the contemporary ideas and ideologies of social value and ranking as reflected by rules of marriage, of garments or processions etc., did not implicate the end to guarantee the function of a social unequal whole which had to stay unequal. With that, the importance of those historical sources depreciates for the purpose of a social order of status and strata. An urban stratification can be ascertained if one succeeds in separating social stratification and social prestige and in replacing the term by empirically detectable and controllable indicators occurring in connection with social prestige in order to prepare general comparability<sup>7</sup>.

As far as the early modern towns are concerned, we assume, following Erich Maschke, that property and wealth defined the degree of social prestige<sup>8</sup>. Therefore, in obtaining the social stratification by the theory of indicators, we interpret the incidence of taxation as a substitutive element suggesting the structure of distribution of property and the structure of social inequality within an urban society as well.

The results of our research on the social and economic history of the town of Kitzingen will be presented in the course of this year. An omnibus volume will contain the design of the socioeconomic stratification, the investigation of property and wealth of the citizenship, an analysis of the political and administrative municipal apparatus and a prosopography of the political elite including the relationships, the economic position, the political participation and, as far as possible, the deviant behaviour of individual members of the elite.

The well-known difficulties of any urban history to get in view the socialpsychological aspect of the urban social systems caused a critical gap in the study of the town of Kitzingen, too. This aspect is not treated, but in a way of patchwork, putting together the accidental and isolated news on behavioral patterns, role expectations and attitudes of the members of the elite, hidden in the minute-books. This is for the simple reason of the special nature of the record series investigated: it was the unusual and the changing data that were put down (Bick-Müller: „Veränderungsdaten“), process-produced historical data are scarcely suitable to fill this gap. So we have to repeat the sentence of Leo Schnore with special emphasise for every research on early modern urban social history: „All in all, historical behavioural studies appear to be virtually impossible, at least if rigorous standards of research are demanded.“

<sup>7</sup> Schäfers, B., Sozialstruktur und Wandel der Bundesrepublik Deutschland. Ein Studienbuch zu ihrer Soziologie und Sozialgeschichte, Stuttgart 1976, p. 252.

<sup>8</sup> Maschke, E., Die Unterschichten der mittelalterlichen Städte Deutschlands, in: Maschke, E. and Sydow, J. (eds.), Gesellschaftliche Unterschichten in den südwestdeutschen Städten, Stuttgart 1967, pp. 1–74.

## Aggregated Census Data and the Analysis of Urbanization in Prussia (1816–1939)

This survey is a preliminary report on a section of the project „Urbanization in Prussia (1815–1939)“ which has been supported by grants of the Deutsche Forschungsgemeinschaft (DFG). For this section the data are now available and analysis has begun. This report deals only with the sources, the data, and the techniques of analysis.

### 1. Subject of the Project

The subject of this project is urbanization as one main branch of the process of modernization. Some of the aspects of urban and social transformations are analyzed as to their temporal and spatial structure. The first aspect is to be investigated by the analysis of the towns and cities of all Prussia. The second point is treated in the framework of three case studies dealing with the changes in the particular town itself and in the surrounding villages. The cases are Mönchengladbach (cotton industry), Saarbrücken (heavy industry), and Breslau (center of commerce). The survey presented here deals exclusively with the analysis of all Prussian towns and cities.

The central hypothesis of the project is that important changes occurring in the transformation of modern industrial society originate from towns and cities and have spread throughout the society. The time-span chosen implies the assumption that, in the case of Germany at least, there is at the beginning, to a certain degree, still a „traditional“ pattern and at the end a „modern“ pattern of urban and social development. Proceeding from demographic changes, this project tries to investigate the causes and consequences of these changes. Three aspects of transformation are analyzed: the spread of urban „innovations“ (sewerage, gas, transportation system etc.), legal and institutional transformations, and longterm structural changes, which are quantitatively identifiable. The report concentrates on this latter point.

## 2. The Object of Investigation

Town/city (Stadt) and urban (städtisch) are the basic criteria for the analysis of urbanization. One crucial point of this analysis, however, is the question of what has to be regarded as town/city or urban. Whereas it is quite clear that the notion of town or city dating from the middle ages is obviously inadequate, it has not yet been replaced by another, commonly acknowledged one. Therefore, instead of beginning with a definition of what is urban, the topic of the investigation is to find quantifiable differences between „town“ and „country“. On this basis it has to be decided whether, where and when there exist fairly recognizable distinctions between towns/cities, or some categories of them, and the country or whether there is only a continuum. This question is to be decided on the basis of criteria developed in the discussion on modernization (e. g. birth rates, mobility, household structure, bureaucratization, budget expenditures, occupational categories).

Concerning the data available for this investigation, there are two different notions of „urban“: one is „town“ as a legal concept as it was used by the Prussian statistics, the other „urban“ in the statistical sense used by the statistics of the German Reich since 1871. Neither of these two criteria is wholly satisfying: the former includes a number of very small communities owing their title to their importance in previous centuries; on the other side the latter comprises a constantly growing number of villages, among them many relatively unimportant communities, so that one might doubt whether this criterion still makes sense at the end of the 19th century. The advantage of the legal definition consists in the fact that it comprises a relatively constant number of communities which is altered only by deliberations and decisions of the government or the crown, whereas the statistical definition allows comparisons on the basis of a firm classification.

## 3. Sources and Materials

The quantitative analysis is mainly based on the census reports of the Prussian statistics („Preußische Statistik“ and preceding statistical works), the statistical series of the German Empire (especially Statistik des Deutschen Reichs, Vierteljahreshefte zur Statistik des Deutschen Reichs) and historical statistics, available now for some regions of Germany, above all concerning demographic data. Moreover, additional statistical sources have been utilized: other statistical publications of Prussia and the German Empire, the statistics and annals for individual towns,



especially cities, and from 1890 onwards, the statistical annals of German towns<sup>1</sup>. Finally, there are many contemporary publications and materials from the archives. The main problem of this material lies in its inconsistency. The level of aggregation is quite different: all towns or all communities of more than 2000 inhabitants, the big cities, the „Stadtkreise“, the towns of more than 10 000 or 5000 inhabitants are examples of totally different levels of aggregation. Besides there are data for every town/community, sometimes only down to a certain size category. The most favorable situation exists for the big cities (more than 100 000 inhabitants). The aggregation on the level of the local community is the main basis of the study. Data concerning intra-urban units are not utilized.

Another problem is brought about by the fact that the methods and criteria differ from census to census so that there are serious obstacles to constructing time series for all the factors being considered. For this project it has been possible to construct a complete time series only for the demographic development. Alongside the importance of population growth for urbanization this is the main reason why the data on population have become the basis of all extended analysis.

Wherever data concerning urban population are available for the whole territory of Prussia with differentiations concerning size, classes or regions (Provinzen, Regierungsbezirke), they have been utilized to trace the general structure of the demographic aspect of the urbanization process. But as these aggregations comprise an ever changing number and an average of very different towns and communities, it seemed necessary to make the individual town/community the starting point for further investigation. On the level of the Prussian state aggregated data on the population of towns with differentiation as to size and region are reported only up to 1910 (sometimes 1925), whereas the first information on communities of more than 2000 inhabitants dates from 1867 and continues to 1939.

Up to 1925 the number of the population of every town is listed for every census which was held since 1816 in intervals of three years, after 1867 intervals of four, and after 1875 five years and since the first world war in irregular intervals<sup>2</sup>. There are lacunas for the years 1822, 1825, 1828, 1834 for which data may be found only for some kinds of towns or regions but which seemed dispensable for this study. Also the information on communities of more than 2000 inhabitants is

<sup>1</sup> „Statistisches Jahrbuch deutscher Städte“; from 1934: „... deutscher Gemeinden“; Silbergleit, Heinrich, Preußens Städte. Denkschrift zum 100jährigen Jubiläum der Städteordnung vom 19. November 1908, Berlin 1908, gives also historical statistical data.

<sup>2</sup> 1816: Jahrbuch für die amtliche Statistik des Preußischen Staates, 1. Jg. (1863), pp. 49 ff.; 1819: Beiträge zur Statistik des preußischen Staats, Berlin 1921, pp. 51 ff.; 1831: J. G. Hoffmann, Neueste Übersicht der Bodenfläche, der Bevölkerung und des Viehstandes, Berlin 1833, pp. 9 ff.; 1837: ders. die Bevölkerung des Preußischen Staats, Berlin 1939, pp. 225 ff.; 1840–1855: Archiv für Landeskunde der preußischen Monarchie, vol. 3 (1858), pp. 212; Tabellen und amtliche Nachrichten über den Preußischen Staat für das Jahr 1858, Berlin 1860, pp. 25 ff.; 1861: Zeitschrift des Königlichen Preußischen Statistischen Bureaus, 2. Jg. (1862), pp. 249 ff.; 1864 u. 1867, 1875–1910: Preußische Statistik, H. 10, 16<sup>II</sup>, 39, 66, 96, 121, 148, 171<sup>I</sup>, 206, 234; 1925: Statistisches Jahrbuch für den Freistaat Preußen, 24. vol. (1928), pp. 26 ff.

reported listwise up to 1925<sup>3</sup>. For 1871, partly for 1925, for 1933, and for 1939 all data were taken from the official register of communities<sup>4</sup>.

The criteria for recording the population have changed in the course of time. Up to 1864 there were two categories which formed the total population, the civil and the military population; the military population included not only the soldiers but also their families and servants at the garrison. For two census years (1831, 1837) only the civil population is recorded. From 1867 to 1910 the actual or present population (*faktische oder ortsanwesende Bevölkerung*) is counted, later on the resident population (*Wohnbevölkerung*). The differences between these categories are regarded as not so important as to prevent comparisons over time.

Since the beginning of the 19th century the techniques of census recording have improved<sup>5</sup>. The first census with recording lists was held in 1840, and in 1864 the recording was effectuated in all parts of Prussia on the basis of household lists. Nevertheless there remained many problems. Some of them result from the definition of town. Sometimes in the lists of towns we find places which neither possessed the constitutional local government law for towns (*Städteordnung*) nor were members of the provincial diets, two major criteria for the status of town. As the statistical lists were drawn up by the provincial authorities and controlled and published by the Royal Statistical Bureau, they were regarded as the official criterion for this analysis. In addition several places in the western regions of Prussia were members of the provincial diet only in combination with another place and only one figure was given for their population. Nevertheless, in the frame of this study, they are regarded as two towns. Sometimes special suburbs (*Vor- und Neustädte*, especially in the case of Magdeburg) were taken as part of the mother town, sometimes they were recorded as separate towns. Other problems resulted from uncertainties concerning town boundaries: in the first half of the 19th century the Prussian statistics counted only the urban settlement within the former walls as urban; in the provinces

<sup>3</sup> 1867: *Preußische Statistik*, H. 16<sup>II</sup>; 1871: *Jahrbuch für die amtliche Statistik des Preußischen Staates*, Jg. 4,1 (1876), pp. 61 ff.; 1875: *Statistik des Deutschen Reichs*, Bd. 25<sup>II</sup>, pp. 60 ff.; 1880: *Statistik d. Dt. Reichs*, Bd. 57<sup>I</sup>, pp. 30 ff.; 1885: *Stat. d. Dt. Reichs*, NF Bd. 32, pp. 42 ff.; 1890: *Vierteljahreshefte zur Statistik des Deutschen Reichs*, 1, H. 2 (1892), pp. 1 ff.; 1895: *Vierteljahreshefte z. Stat. d. Dt. Reichs*, 6, H.3 (1897), pp. 1 ff.; 1900: *Stat. d. Dt. Reichs*, vol. 151, pp. 736 ff.; 1905: *Vierteljahreshefte z. Stat. d. Dt. Reichs*, 16, H. 4 (1907), pp. 142 ff.; 1910: *Stat. d. Dt. Reichs*, vol. 240, Anhang, pp. 2 ff.

<sup>4</sup> 1871: *Die Gemeinden und Gutsbezirke des Preussischen Staates und ihre Bevölkerung*, 11 issues, Berlin 1873/74 („*Gemeinde-Lexikon*“); 1925: *Gemeinde-Lexikon für den Freistaat Preußen. Nach dem endgültigen Ergebnis der Volkszählung vom 16. Juni 1925*, 14 vols., Berlin 1930–1933; *Vorläufige Ergebnisse der Volkszählung im Deutschen Reich vom 16. Juni 1925* (Sonderheft 2 zu *Wirtschaft und Statistik*); 1933: *Amtliches Gemeindeverzeichnis für das Deutsche Reich auf Grund der Volkszählung 1933*, Berlin 1933 (*Stat. d. Dt. Reichs*, vol. 450); 1939: *Amtliches Gemeindeverzeichnis für das Deutsche Reich auf Grund der Volkszählung 1939*, Berlin 1940 (*Stat. d. Dt. Reichs*, vol. 550).

<sup>5</sup> Blenck, E. (ed.), *Das Königliche Statistische Bureau im ersten Jahrhundert seines Bestehens 1805 bis 1905*, Festschrift, Berlin 1905, pp. 21 ff.

of Rhineland and Westphalia the town was sometimes part of a Bürgermeisterei, a heritage of the French administration. This situation caused some confusion among the local authorities and in the statistics. Others were produced by confusing urban districts (Stadtkreis) or police districts (Polizeibezirk) with town boundaries. Many minor towns (Flecken) are to be found in the official lists of towns, whereas numerous markets (also Flecken) were counted as villages. In 1890 and 1905 official lists of towns were drawn up by the Prussian Statistical Bureau<sup>6</sup>.

The criterion of 2000 inhabitants for urban places in the statistical sense meets with problems, as well. Contrary to the French notion of „agglomeration“ (2000 inhabitants at the principal place of the community)<sup>7</sup>, in Germany the criterion has been related to the whole community. After the occupation of Schleswig-Holstein and other territories in the war of 1866, places are recorded (Kirchspiele) in the list of 1867 which were left aside insofar as they appear only once. The investigation confines itself to Prussia in its respective borders.

#### 4. The Data Base „Urbanization in Prussia“

The data base, collected on code sheets, punched on cards, and stored on tapes, contains 3398 cases and 152 variables. Each case (town/community) comprises 6 cards; 40 variables relate to the 30 census dates (10 variables relate to military population, 4 are empty). The other variables are intended to give information on special aspects of urbanization and to process the population data for cross sectional analysis at different points of time and for longitudinal sections.

The following variables are designed for this purpose:

- affiliation to administrative district (Regierungsbezirk) and changes
- date of bestowal of municipal statutes or the status of town and loss of municipal laws
- membership in the provincial diet
- tax group in 1837
- date of becoming an urban district (Stadtkreis)
- incorporation into another town (community/name of this community)

<sup>6</sup> Verzeichnis derjenigen Gemeinden im preußischen Staate, welche nach Lage der Gesetzgebung als Städte zu behandeln sind, in: Zeitschrift des Königlich Preussischen Statistischen Bureaus, 33 (1890), pp. 51 ff.; Kühnert, F., Die Städte Preußens, in: Zeitschrift des Königlich Preussischen Statistischen Bureaus, 44 (1904), pp. 276 ff.

<sup>7</sup> Horstmann, Kurt, Die Gliederung nach Stadt und Land in der Statistik, in: Allgemeines Statistisches Archiv, 28 (1939), pp. 22 ff.; Löwinger, Istvan Daniel, Stadt und Land in der Statistik (unter besonderer Berücksichtigung der städtischen Bevölkerung in der Schweiz), Diss. Zürich 1970, pp. 27 ff.

- entry in the official town lists of 1880 and 1905
- place of residence of public authorities in the years 1819, 1848, 1885, 1910, 1939 (Staat, Provinz, Regierungsbezirk, Kreis)
- geographical co-ordinates
- for every census date changes concerning the affiliation to the administrative district, the status and – as far as possible – the community district and, if necessary, the reason for leaving the data out (loss of the status of town, falling under the threshold of 2000 inhabitants incorporation, territorial changes) are given.

The data for some of these variables had to be obtained by individual investigation of every case by several means (among other sources the *Deutsches Städtebuch*)<sup>8</sup>.

As the administrative units often do not coincide with zones of urbanization processes it was planned to demonstrate the distribution and the spatial structure of major results by maps. For this purpose two variables were created recording geographic co-ordinates (system Gauss/Krüger). For the actually existing communities in the territory of the Federal Republic the co-ordinates were taken from the official register of communities<sup>9</sup>. For the territories outside the Federal Republic or for those places which have disappeared by incorporation or fusion of communities they were taken from contemporary maps<sup>10</sup>.

<sup>8</sup> Among others: Keyser, Erich (ed.), *Deutsches Städtebuch*, 2 vols., Stuttgart/Berlin 1939, 1941; since 1952 the volumes in the other German regions. Historical statistical surveys: *Gemeindestatistik des Landes Nordrhein-Westfalen. Bevölkerungsentwicklung 1871–1961*, Düsseldorf 1964 (Beiträge zur Statistik des Landes Nordrhein-Westfalen, Sonderreihe Volkszählung 1961, Heft 3 c); *Historisches Gemeindeverzeichnis für Hessen*, 2 vols., Wiesbaden (1968); *Beiträge zur historischen Statistik Schleswig-Holsteins*, Kiel 1967; Uelschen, Gustav, *Die Bevölkerung in Niedersachsen 1821–1961*, Hannover 1966. On the date of origin also: Haase, Carl, *Die Entstehung der westfälischen Städte*, 2nd. ed., Münster/Westf. 1965; Weczerka, Hugo, *Stadt- und Marktgründungen und Stadtabgänge in Schlesien 1450–1800*, in: *Zeitschrift für Ostforschung*, 23 (1974), pp. 193–261.

<sup>9</sup> *Amtliches Gemeindeverzeichnis für die Bundesrepublik Deutschland – Ausgabe 1957* –, Stuttgart 1958, pp. 565 ff.

<sup>10</sup> Especially: *Topographische Karte* (1:25 000, Meßtischblatt, since 1875); *Topographische Karte* (1:100 000, Generalstabskarte, since 1878).

## 5. Processing the Data

The processing of the data pursues various aims:

- a) The completion of the data at the aggregation level of Prussia
- b) Processing of variables concerning special aspects of urbanization (occurrence of new towns, loss of status etc.)
- c) Calculation of growth rates of the population
- d) Analysis of the factors of urbanization processes and their consequences by means of population growth rates and selected data on other aspects
- e) Demonstrating the spatial and regional structure and patterns of urbanization processes in different regions of Prussia by means of maps

For item a above, missing data for the whole population of towns for the census dates 1925, 1933, 1939 are completed by counting all cases in the sample and their values (i. e. the urban population) differentiated according to size classes and administrative districts (Regierungsbezirke), (see Table 1).

For item b, the untransformed data on population serve a basis for inquiring into the changes of rank-order for Prussia and for individual administrative districts (Regierungsbezirke). The frequency distribution of some phenomena differentiated as to date, size, and region allows statements on changes of the urban system (Table 2): new towns/communities of more than 2000 inhabitants, loss of towns/communities by incorporation or fusion, new urban districts (Stadtkreise). The results are discussed with regard to periods and zones of heavy or moderate changes.

For item b, the untransformed data on population serve as basis for inquiring into the distribution over several growth classes (Tables 3a and 3b) allows insights into the structure of urban growth as to different periods. The processing of growth rates for population size and other variables gives information on whether growth depends on initial population size, on the age of the town, on its formal status of town, on its administrative functions, on its geographic situation, or on its industrial activity (tax class 1837).

Furthermore these growth rates are the starting point for the analysis of causes and consequences of urban growth (item d above). As there are no data for all cases over time and as this aim would go beyond the scope and the possibilities of this project all the cases of the extreme groups (decrease and extreme growth) and samples of the other groups are investigated by statistical tests in so far as data are available. The basis of this investigation are lists of growth rates.

Some of the results are to be plotted by a program that was originally developed for small scale regions and relatively few cases<sup>11</sup>. It enables one to locate a place by means of two co-ordinates, to use different symbols, to vary their size, and permits

<sup>11</sup> Kern, Hans, and Schulz, Karin, THEKAR — Ein Programm zur automatisierten Herstellung thematischer Karten nach dem Diagrammprinzip, in: Kartographische Nachrichten, 26 (1976), pp. 94 ff.

Table 1: Urbanization in Prussia (legal concept of „town“)

	Number of Towns	Population Living in Towns	Military Population Living in Towns	Share of Urban Population in % (including Military Population)
1816	1,020	2,731,439	150,094	27.9
1819	1,027	3,023,292 <sup>2</sup>		27.5
1822	1,041	2,997,973	169,960	27.2
1825		3,223,173 <sup>2</sup>		26.3
1828		3,367,433 <sup>2</sup>		26.5
1831	983	3,334,140	258,215	27.6
1834		3,684,671 <sup>2</sup>		27.3
1837	972	3,639,983	214,513	27.3
1840	976	3,861,019	205,247	27.2
1843	979	4,060,030	202,678	27.6
1846	980	4,308,208	200,740	28.0
1849	980	4,324,813	257,385	28.1
1852	988	4,638,511	176,698	28.4
1855	993	4,760,728	207,428	28.9
1858	994	5,040,033	195,966	29.6
1861	1,000	5,362,141	263,711	30.4
1864	1,001	5,741,982	274,285	31.2
1867	1,272	7,164,638	278,724	31.1
1871	1,290	8,000,931		32.5
1875	1,288	8,791,834		34.2
1880	1,287	9,707,802		35.6
1885	1,287	10,554,596		37.3
1890	1,263	11,786,061		39.3
1895	1,266	12,954,591		40.7
1900	1,266	14,847,846		43.1
1905	1,279	16,866,963		45.2
1910	1,276	18,963,785		47.2
1925 <sup>1</sup>	1,099	20,854,855		54.7
1933 <sup>1</sup>	1,098	23,494,292		58.8
1939	1,152	25,627,948		61.5

<sup>1</sup> Without Saarland

<sup>2</sup> Whole population (including military population)

Sources: 1819, 1837, 1846, 1852, 1855, 1861–1867, 1875–1910: see foot note 2;  
 1816, 1822, 1831, 1840, 1849, 1858: Das Anwachsen der Bevölkerung im Preußischen State seit 1816, in: Zeitschrift des Königlich Preußischen Statistischen Bureaus, 1. Jg. (1861), S. 9 ff.; 1825, 1828, 1834: Preußische Statistik, H. 188, Teil B, S. 2 f.; 1843: W. Dieterici, Die Statistischen Tabellen des preußischen Staates nach der amtlichen Aufnahme des Jahres 1843, Berlin 1845, S. 242; 1871: R. Jannasch: Das Wachstum und die Concentration der Bevölkerung des Preußischen Staates, in: Zeitschrift des Königlich Preußischen Statistischen Bureaus, 18. Jg. (1878), S. 263 ff.; 1925, 1933, 1939: Results of data base „Urbanisierung in Preußen“.

**Table 2: Inclusion and Exclusion of Prussian Towns into the Data Base**  
(differentiated into Prussian provinces)

	1816-42		1843-66		1867-89		1890-1918		1919-39		1816-1939		fluctu- ating
	incl.	excl.	incl.	excl.	incl.	excl.	incl.	excl.	incl.	excl.	incl.	excl.	
Ostpreußen	—	—	—	—	—	—	—	—	1	—	1	—	—
Westpreußen	—	2	—	—	—	1	1	1	—	—	2	3	1
Posen	—	2	—	1	—	11	—	4	—	—	—	18	—
Brandenburg	—	—	1	1	—	2	4	—	4	1	9	4	2
Pommern	—	—	1	—	—	—	—	1	—	—	1	1	—
Schlesien	1	1	3	3	3	—	2	1	5	1	13	6	3
Sachsen	—	5	3	—	1	—	—	—	4	—	8	5	6
Westfalen	5	35	2	—	3	—	5	1	36	—	51	36	—
Rheinprovinz	33	39	4	—	5	1	3	—	15	—	59	40	11
Schleswig- Holstein					1	—	3	—	—	—	4	—	—
Hannover					1	—	3	—	7	1	11	1	—
Hessen-Nassau					4	5	—	—	7	—	11	5	—
Hohenzollern- Sigmaringen					—	5	—	—	—	—	—	5	—
all together	40	84	14	5	19	24	21	8	79	3	173	124	23

Table 3a: Growth of Prussian Towns (annual growth rate in %)

	1816-40	1840-71	1871-90	1890-1910	1910-39	1816-1939
decline ( $< 0,0\%$ )	19 2.1%	81 8.5%	289 23.3%	253 20.4%	88 8.4%	10 1.4%
small growth rate ( $0,0-0,6\%$ )	82 8.9%	264 27.6%	334 26.9%	275 22.1%	298 28.4%	190 26.1%
( $0,6-1,2\%$ )	229 24.9%	304 31.8%	267 21.5%	232 18.7%	340 32.4%	308 42.3%
strong growth rate ( $1,2-1,8\%$ )	318 34.6%	178 18.6%	134 10.8%	206 16.6%	181 17.2%	146 20.1%
( $1,8-2,4\%$ )	169 18.4%	56 5.9%	84 6.8%	104 8.4%	74 7.0%	48 6.6%
extremely strong growth rate ( $> 2,4\%$ )	102 11.1%	68 7.1%	132 10.6%	173 13.9%	70 6.7%	26 3.6%
all together	919 100%	957 100%	1,240 100 %	1,243 100%	1,051 100%	728 100%
mean	1.46%	0.98%	0.84%	1.09%	0.96%	1.00%
max.	5.46%	7.68%	7.48%	10.67%	7.39%	4.11%
min.	-5.83%	-2.82%	-3.05%	-1.64%	-0.43%	-0.48%



**Table 3b: Growth of Rural Communities with more than 2000 Inhabitants**  
(annual growth rate in %)

	1871-90	1890-1910	1910-39	1871-1939
<b>decline</b> ( < 0,0%)	80 16.5%	57 8.5%	95 13.2%	37 11.6%
<b>small growth rate</b> (0,0-0,6%)	124 25.5%	121 18.1%	189 26.2%	105 32.8%
(0,6-1,2%)	81 16.7%	105 15.7%	221 30.7%	91 28.4%
<b>strong growth rate</b> (1,2-1,8%)	57 11.7%	118 11.8%	110 15.3%	48 15.0%
(1,8-2,4%)	39 8.0%	66 9.9%	51 7.1%	18 5.6%
<b>extremely strong</b> <b>growth rate</b> ( > 2,4%)	105 21.6%	240 35.9%	55 7.6%	21 6.6%
<b>all together</b>	486 100%	668 100%	721 100%	320 100%
<b>mean</b>	1.40%	2.20%	0.99%	0.91%
<b>max.</b>	11.01%	17.19%	9.18%	5.60%
<b>min.</b>	-3.09%	-3.02%	-3.67%	-0.86%

the overlapping of symbols. The program has been adjusted to the tasks of this project, but some further development is necessary. As the system Gauss/Krüger contains some cuts in the grid system, a plot of the six strip which cover the territory of Prussia has to be assembled with the maps which are available for the different locations of Prussia (1815–1866/1867–1918/1919–1933). On the other hand, the geographic co-ordinates (measured in kilometers) may serve for analysis on the structure of the urban system. Thus by calculating the distances between different types/sizes of towns, typical patterns of spatial distribution may be investigated.

## 6. Problems and Experiences

Some problems concerning the processing arise because the data are destined both for cross sectional and cross temporal analysis and because at every census the relation of a case to a status group (village, town, urban district etc.), or to an administrative district or other conditions may change. These definitions, the task definitions, and the number of cases cause a high demand for workspace. Up to now the data have been processed by SPSS, Version 6.0, or by special FORTRAN-Programs in the Wissenschaftliches Großrechenzentrum Berlin (GRZ).

As the SPSS system needs much space, too, it was possible to process the data only late in the evening or at night. Also concerning other aspects the SPSS program proved not well suited for the tasks in question: e. g. the summing up of the values of all cases for Prussia and the size groups was done by FORTRAN.

To meet these problems, tests are intended with the program DIANA developed at the Zentralinstitut für sozialwissenschaftliche Forschung (FU Berlin) implemented at the GRZ Berlin and the University of Hamburg<sup>12</sup>. It is oriented to interactive applications of documentation, retrieval and numerical analysis. The advantages of this program consist in its ability to handle large quantities of data in its relatively simple instruction language, in its small requirements for workspace, and its quicker processing.

<sup>12</sup> Tiemann, Rainer, DIANA-Programmbeschreibungen (RAPROSYS/DIANA), Computer-Druck, Berlin 1977 (Änderungsstand v. 21. 6. 1977).



### III. Analysis of Life Histories



Life Histories:  
Problems and Prospects of Longitudinal Designs

The aim of the paper is to provide a general overview of problems and prospects of life history studies based on longitudinal designs and interviews as a data source. The value of life history studies is hereby seen from two angles: as an attempt to unravel an individual's development over time in the context of changing roles, opportunities and constraints and as an attempt to unravel the influence of specific historical conditions on an individual's life course. Longitudinal designs are for the following defined in the broadest sense, as studies which encompass a time perspective, tracing units of observations — such as individuals or aggregates of individuals — over time.

The paper will restrict itself to a discussion of methodological strategies and possibilities of the various longitudinal designs. It will omit statistical and technical details associated with the analysis of the data themselves<sup>1</sup>. It hopes to sensitize the reader to the possibilities of bringing time back into sociological analysis and to illustrate the prospects and the problems which are linked with such endeavours. Three basic kinds of longitudinal studies will be discussed: cohort studies, prospective longitudinal studies and retrospective studies on the basis of retrospective questions. Trend studies will not be specifically dealt with in this article, since they do not themselves necessarily incorporate a life history perspective. However, cohort studies are usually based on trend studies and many of their problems are consequently identical. Our discussion of the problems of cohort studies has therefore implications for the study of trends as well.

<sup>1</sup> For recent trends in the development of the methodological literature pertaining to data collection and analytical techniques in general and a critique of the neglect of data quality problems see Scheuch, Erwin K., *Forschungstechniken als Teil der Soziologie heute*, in: Lepsius, M. R. (ed.), *Zwischenbilanz der Soziologie heute*, Stuttgart 1976, pp. 110 passim.

## 1. Cohorts, Prospective and Retrospective Longitudinal Studies: a Note on Terminology

Before going into details, a clarification of terms is in order. The need for clarification is the greater since the literature is replete with different terminologies and connotations for the same kind of longitudinal designs<sup>2</sup>. The greatest agreement, without doubt, can be found with regard to the term „trend study“. A trend study is usually conceived as a study of the same population, though not necessarily the same individuals, measured at different times with regard to some common variables. Repeated cross-sectional measurements of the population with regard to voting intentions, for instance, would constitute a trend study in this sense.

The terminological agreement is less concerning *cohort* studies. Cohort studies in the broadest sense might be defined as studies of cohorts, i. e. of aggregates of people (within some population) who experienced the same event within the same time interval. Usually birth has been taken as the defining event, but other events, such as entry into an institution within a limited period of time, could be used equally well. When the word cohort is used without modifier, a birth cohort is generally meant<sup>3</sup>. Many methodological designs are compatible with this kind of conception. Thus, some authors refer to a cohort analysis when different cohorts within a cross sectional survey are compared with each other<sup>4</sup>. Since cohort is identical with age in this case, the true delineation of cohort effects is not possible. Other authors conceive of cohort studies as studies which measure the same individuals of a given cohort at different times (in a panel study for example)<sup>5</sup>. A further version is found among those who use retrospective interviewing in order to reconstruct the past. If the analysis is then based on a differentiation of a cross sectional survey according to cohort or age and the present data are compared with the reconstructed ones on a cohort basis, the term cohort analysis has been applied

<sup>2</sup> For divergencies in terminological usage see for instance Babby, Earl R., *Survey Research Methods*, Belmont/Cal. 1973, pp. 62 passim; Galtung, Johan, *Theory and Method of Social Research*, London 1967, pp. 84 passim; Goldfarb, Nathan, *An Introduction to Longitudinal Statistical Analysis. The Method of Repeated Observations*, Glencoe/Ill. 1960; Wall, W. D., and Williams, H. L., *Longitudinal Studies and the Social Sciences*, London 1970; Nehnevasja, Jiri, *Analyse von Panel-Befragungen*, in: König, René (ed.), *Handbuch der empirischen Sozialforschung*, Vol. 2, Part 1, Stuttgart 1973, pp. 192 passim; Davies, James A., *Panel Analysis. Techniques and concepts in the interpretation of repeated measurements*, mimeo, Chicago 1963.

<sup>3</sup> See for instance Glenn, Norval D., *Cohort Analysis*, Beverly Hills and London 1977; pp. 8 passim; Ryder, Norman B., *The Cohort as a Concept in the Study of Social Change*, in: *American Sociological Review*, 30 (1965), pp. 843–861.

<sup>4</sup> See for example Butler, David, and Stokes, Donald, *Political Change in Britain*; Harmondsworth 1971, pp. 78 passim.

<sup>5</sup> See the literature quoted in: Wall, *Longitudinal Studies*.

there as well<sup>6</sup>. Finally there is a version, mostly favoured by those engaging in secondary analysis of surveys. According to this version, cohort studies are studies which compare *different* cohorts as *aggregates* on at least some *common variables* at *different* points in time. The aggregate might hereby consist of the same individuals as in former surveys (census data for example) or they might represent only a sample of the same population and thus not necessarily the same individuals (as in trend surveys)<sup>7</sup>. In the following, we want to use the term cohort studies in this sense. So called cohort studies measuring the same individuals of a given cohort over time on an individual level will be treated as a specific case of prospective longitudinal studies. And so called cohort studies based on retrospective questioning will be referred to as a special case of retrospective studies.

Prospective *longitudinal* studies are studies which involve repeated measurement of the same individuals over time. Panel analyses are an example. Experiments, in which a particular „treatment“ is administered to a portion of the study group between repeated measures of the same subjects, could in principle be seen as of longitudinal design as well. However, as long as the before and after design involves a short time period only, we prefer not to include experiments. Prospective longitudinal studies might take on different forms according to their variables: If the same individuals are repeatedly measured using the *same* variables the term panel analysis is usually applied. If the same individual is repeatedly measured using *different* variables, the term prediction studies is not uncommon. Both kinds of longitudinal analysis usually mix, since even in prediction studies, concerning marital adjustment or delinquency for instance, some variables remain the same in the subsequent measurement waves. Probably for this reason, many authors see the terms panel and longitudinal study as equivalent<sup>8</sup>.

Prospective longitudinal studies might also take on different forms with regard to the time dimension of sampling. One could start from data collected in the past by other researchers and follow them up. One could, of course, also start from the present and proceed to the past for given individuals by taking recourse to archival data. The latter kind of studies have sometimes been called retrospective studies<sup>9</sup>. However, sampling from the past to the present or from the present backwards or from the present to the future does not alter the basic longitudinal design of obtaining repeated measurements on the same individual. Some studies, moreover, have

<sup>6</sup> E. g. Lane, Angela, The Occupational Achievement Process 1940–1949: A Cohort Analysis, in: American Sociological Review, 40 (1975), pp. 472–482.

<sup>7</sup> Before the use of surveys for cohort studies, only census data were used. This is still reflected in J. A. Davies' definition from 1963, where the comparison of cohorts is conceived as a comparison of the same individuals. See Davies, Panel Analysis.

<sup>8</sup> Some authors on the other hand have seen the difference between panel and longitudinal studies in the time dimension: panel studies involving a shorter time span.

<sup>9</sup> See for instance Goldfarb, Introduction, p. 41; O'Donnell, J. A., The Methodology of Retrospective Studies, paper presented at the Conference on the Epidemiology of Drug Dependence, London 1972.



combined a number of approaches of this sort. For instance in the Danish-Swedish Project Metropolitan which consists of a longitudinal study of youth born in 1960 and which began in 1964, data about child delivery were combined with data collected in a prospective longitudinal design<sup>10</sup>. We would therefore prefer to use the term retrospective study for those designs which use retrospective questions in order to reconstruct the life history of individuals.

In *retrospective* or quasi-longitudinal studies the subjects have only one measurement in time but data pertaining to different points in time; individuals are asked to recollect the events, decisions and definitions they brought to the situation in past times. The longitudinal approach in this design is thus one based on the memory of the respondents. Various data collection strategies have been associated with this approach, ranging from more or less unstructured forms of data collection (such as in qualitative interviewing or analysis of specially elicited written autobiographies) to a highly structured, closed interviewing. Samples have accordingly varied between small samples of special social groupings to large samples of the population in general. Finally, retrospective studies have been performed on single surveys or — in a few cases — on trend studies<sup>11</sup>.

## 2. Cohort Studies

### 2.1 Characteristics

The cohort concept, though rather recent in the social sciences, is not entirely new to the literature. Under the label „generation“ it has had some tradition, especially among historians of the art, and foremost in Karl Mannheim's classical treatise on the problem of generations. The other field of application has been in the field of demography where it has been used mainly in the study of fertility<sup>12</sup>.

<sup>10</sup> Janson, Carl-Gunnar, Project Metropolitan. A presentation, in: Project Metropolitan, Research Report No. 1, Stockholm 1975.

<sup>11</sup> Using trend studies and retrospective questions might be useful in delineating continuous trends in age-limited populations (such as youth) even where the surveys have not been done on a continuous basis. In this case the cumulation of samples allows a cumulation of individuals from different surveys who have made certain experiences in certain years. The limitations brought into the design by having age limits are thus compensated and a description of yearly trends becomes possible. For such a strategy see for example Reuband, Karl-Heinz, Eine neue Drogengeneration? Zur Analyse sozialen Wandels im Bereich abweichenden Verhaltens, in: *Angewandte Sozialforschung*, No. 5/6 (1978), pp. 321–329.

<sup>12</sup> See Pfeil, Elisabeth, Der Kohortenansatz in der Soziologie, in: *Kölner Zeitschrift für Soziologie*, 19 (1967), pp. 645–657; Mannheim, Karl, Das Problem der Generationen, in: v. Friedeburg, Ludwig (ed.), *Jugend in der modernen Gesellschaft*, Köln and Berlin 1965, pp. 23–48.

For both traditions, the crucial aspect of cohorts is the notion that historical influences impinge on given birth cohorts and lead to specific generational effects. „The members of any cohort are entitled to participate in only one slice of life — their unique location in the stream of history. Because it embodies a temporally specific version of the heritage, each cohort is differentiated from the others“<sup>13</sup>. The most common approach is to conceive the period of youth as the period where generation formation takes place. It is in youth where, as Mannheim says, experiences have a tendency to form a „natürliches Weltbild“, a taken for granted world view, which serves as the basic layer of accumulated experience. It is the foundation which acts as a filter, structuring subsequent experiences<sup>14</sup>.

The value of the cohort approach is usually seen both from a macrosociological perspective and from a life history perspective. The macrosociological point of view refers to „demographic metabolism“ to the structural transformation caused by the coming and going of generationally differentiated cohorts within a population system<sup>15</sup>. Social change, according to this perspective, might be simply due to an exchange of differently influenced and orientated generations. The life-history perspective focuses on the historical and developmental career of the individual. The cohort record is here seen as a *macro-biography*, as the aggregate analogue to the individual life history. By comparing the life courses of different cohorts having undergone different historical experiences, an attempt could be made to separate historically generated generational life histories and history-independent „natural“ courses of individuals' lives, attitudes and behaviors. Although the first systematic presentation of those concepts was given to social scientists at the end of the fifties, it was not until the late 1960's and especially the 1970's that cohort analysis received increasing attention and actually lead to empirical investigations. The reason for this delay is probably due to the rather late introduction of the concept into the social science literature on the one hand, and the lack of available appropriate surveys on the other. With the passage of time and the accompanying enlargement of data archiving, surveys became increasingly available which allowed cohort analyses. It is not unlikely furthermore, that social changes happening in most western countries in the late sixties produced more sensitivity to generational phenomena. It is at least quite noteworthy that the development of student unrest and various forms of countercultures brought a revival in generational thinking within the sociological literature. Studies on youth and countercultures and youth and political rebellion quite often now use the term generation with all its theoretical implications<sup>16</sup>.

The usual approach in the study of cohorts, as already mentioned, is to take a number of different surveys collected at different points in time and to divide them into birth cohorts. Simply using a cross-sectional survey is not sufficient, since its

<sup>13</sup> Ryder, Cohort, p. 844.

<sup>14</sup> Mannheim, Generationen, p. 41;

<sup>15</sup> Ryder, Cohort, p. 843.

<sup>16</sup> See for instance The Journal of Social Issues, 30 (1974), No. 3 and No. 4.

age distribution does not allow separating aging and cohort or „generational“ effects: in this case, age interpreted as cohort, is perfectly correlated with chronological age for all observations. If we have a concentration of conservative attitudes in the older population, for instance, we do not know whether this finding is due to aging or to the impact of the historical period of time in which the individuals were brought up.

Incorporating additional cross-sectional surveys into the design offers a solution when the other surveys have been collected at different points in time and permit a comparison on a cohort basis. Birth age and chronological age become decorrelated under these conditions, thereby providing the two variables to distinguish aging and cohort effects. We can, for instance, now compare the 20–29 years old in 1950 with the 30 to 39 years old in 1960 and treat both as being representative of the same cohort (see Table 1). Such a procedure can only be chosen, of course, if the cross-sectional surveys refer to the same population. This population must be closed: the rate of compositional change due to migration must be rather low. Surveys within a community or a regionally restricted area rather than within the nation therefore usually fall out of this realm: the rate of exchange due to geographical mobility might be too high<sup>17</sup>.

*Table 1: Illustration of the Logic of Cohort Analysis*

Cohorts	Survey Conducted in			← Age
	1950	1960	1970	
A	20–29	20–29	20–29	← Age
B	30–39	30–39	30–39	
C	40–49	40–49	40–49	
D	50–59	50–59	50–59	
	60–69	60–69	60–69	

<sup>17</sup> It has been proposed to control for place of birth and length of residence in order to check possible biases due to analyzing cohorts on a community basis. See Duncan, Otis D., *Measuring Social Change via Replication of Surveys*, in: Land, Kenneth C., and Spilerman, Seymour (eds.), *Social indicator models*, New York 1975, p. 117. However, migration is selective and we don't know very much about those emigrating. For this reason cohort studies on a community basis

## 2.2 Aging, Cohort- and Period Effects

Inspection of the cohort table usually permits an idea about the relevance of aging and cohort effects. An *aging* effect is said to exist when comparisons within cohorts (intracohort comparisons) — as indicated by diagonal reading of the table from the upper left to the lower right — show systematic patterns of change irrespective of the cohorts. Pure aging effects exist when each set of cross-sectional data is identical and when intracohort comparisons show the same amount of change with increasing age as the equivalent movements in the cross sectional sections (see Table 2a).

*Cohort* effects are said to exist when there is an „early imprinting which, while it may be overlaid by other kind of effects, including life cycle ones, during the life span, will always leave its characteristic mark as a deviation of the generation affected from neighboring generational cohorts“<sup>18</sup>. Pure cohort effects exist when the cohorts keep their peculiar characteristic in the process of aging and when accordingly no variations exist in the cohort diagonals (see Table 2b).

The third class of effects to be mentioned are *period* effects. Period effects are said to exist when all age and cohort groups are (temporarily) affected by common historical experiences. In cases of pure period effects, as shown in table 2c, changes in all age and cohort categories are the same.

Table 2: Cohort Table Showing Hypothetical Data (Percentage) in Which All Variations is due to:

	a) Aging Effects			b) Cohort Effects			c) Period Effects		
Age	1950	1960	1970	1950	1960	1970	1950	1960	1970
20–29	30	30	30	30	20	10	50	40	30
30–39	40	40	40	40	30	20	50	40	30
40–49	50	50	50	50	40	30	50	40	30
50–59	60	60	60	60	50	40	50	40	30
60–69	70	70	70	70	60	50	50	40	30
Total*	50	50	50	50	40	30	50	40	30

\* Standardized to an age distribution with an equal number of persons at each age level.

usually remain problematic. The only exception are those studies which refer to populations with a low rate of migration in the period of observation, such as youth in schools. In these cases problems of migration might not be high.

<sup>18</sup> Converse, Philip E., *The Dynamics of Party Support*. Cohortanalyzing party identification, Beverly Hills and London 1976, p. 74.

In many cases, however, period effects do not operate as simply as is often assumed. They might have different effects on different age and cohort groupings and thus make it problematic to delineate the exact impact of period influences. The launching of a military draft (e. g. during the Vietnam war in the U.S.A.), for instance, affects different age groups than a breakdown in pension schemes. The same kind of event might even have contrary effects in different age groupings. Period effects (as well as the other effects) might hence be interactive and not additive. Besides it should be noted that period effects, when not being rather temporary, might have long lasting effects and thus become generation forming<sup>19</sup>.

Period effects have often not been given adequate attention at the beginning of research on cohorts. Studying aging and cohort effects without attention to period effects, however, could have serious consequences. If we compare conservatism in different age and cohort groupings, for instance, and find increasing conservatism with age on the cross-sectional level but a decrease with age on the cohort level, this result must not necessarily mean that there are no aging effects leading to stronger conservatism. The results in this case might have been obtained by the influence of period effects working towards a decrease of conservatism in age and cohort groupings<sup>20</sup>.

The biggest problem of cohort studies, not fully recognized until recently, is the problem of *underidentification*. What we have are three classes of effects — age, cohort and period effects — and only two diagnostic variables. We obtain the cohort effects, for instance, as an interaction between age and period. We might conversely obtain age as an interaction between cohort and period and obtain period as an interaction between age and cohort. The third effect is hence only found by recourse to the other two effects. In every case, the independent variable to be used for the explanation is a *perfect* function of the other two variables which are to be used as independent or control variables. This confounding of variables makes it technically impossible to delineate the exact influences of the three effects. The table presented as Table 2b for instance showing pure cohort effects could theoretically also have been produced by a mix of aging and period effects each offsetting each other: If an aging effect would have taken place *and* a period effect which caused a ten percent drop in every age and cohort between 1950 and 1960 we would have obtained the same distribution for 1950 and 1960 as in Table 2b. The problem of identification is „indeterminate, at least within the immediate terms of conventional proof through the manipulation of numerical matrices of cohort values, however long the time span of observation may be“<sup>21</sup>. A solution to the problem, as

<sup>19</sup> See Glenn, Norval D., *Cohort Analysis*; Converse, Dynamics.

<sup>20</sup> The gradual growing awareness of the different effects over time can be seen in the discussion of the Crittenden article. See Crittenden, John, *Aging and Party Identification*, in: *Public Opinion Quarterly*, 25 (1962), pp. 648–657; Cutler, Neal E., *Generation, Maturation, and Party affiliation: a Cohort Analysis*, in: *Public Opinion Quarterly*, 33 (1970), pp. 583–588; Crittenden, John, *Reply to Cutler*, in: *Public Opinion Quarterly*, 33 (1970), pp. 589–591.

<sup>21</sup> Converse, P. E., *Party Support*, pp. 17 *passim*. For the problem of underidentification see also Glenn, *Cohort Analysis*, pp. 13 *passim*.

suggested by Philip Converse and Norval Glenn, is to take „side information“ into consideration. Side information might permit at least one of the effects to be ruled out as an influential factor. This ruling out, of course, very often is of a tentative nature. It rests on a certain amount of plausibility. If there is a sharp drop in the number of adherents to a certain attitudinal position in the population within a short span of time, there is good reason to believe that this drop cannot have taken place by an exchange of generations. Period effects must have been operating. It seems reasonable therefore, as Philip Converse has done in one study, to perform the analyses separately for the periods which seem stable and unstable<sup>22</sup>.

On the basis of „side information“, it often seems possible to come to some conclusions about the impacts of different effects. In some cases it might also be sufficient simply to have some information about the attitudinal or behavioral state of an aggregate regardless of knowing the exact nature of the effects: Seen from a life history point of view, age, cohort and period effects *together* might constitute an important experience in the life course, whatever the role of specific effects is. Choosing a cohort approach also often allows some inferences about changes on the societal level: we might infer from the analysis whether a perceived change, for instance the complete reversal in the relationship between education and capital punishment in W. Germany between 1950 and 1970, is due to specific cohorts only or to a general intracohort change<sup>23</sup>. The dependent variables might hereby, as the example suggests, not only refer to percentage distributions, but also to relationship between variables, such as sex, education and some attitudinal variable or between attitudinal variables themselves<sup>24</sup>.

Cohort approaches might be performed on birth cohorts in general or on subgroups within birth cohorts, for instance on sex and educational groupings. An important presupposition for such a procedure, of course, is that there is little or no movement from one category to the other, as is the case with sex or education. Occupational groupings seem to be less apt for such a procedure, since there is a relatively high amount of intragenerational mobility over the life span, making a strict cohort comparison impossible. In total, then, there are many cases where the cohort approach seems to be quite fruitful for an understanding of historical influences and the life history of individuals.

<sup>22</sup> Converse, Party Support.

<sup>23</sup> See Reuband, Karl-Heinz, Die Einstellung zur Todesstrafe im Wandel, in: Steinert, Heinz (ed.), Der Prozeß der Kriminalisierung. Untersuchungen zur Kriminalsoziologie, München 1973, pp. 51–61; Reuband, Karl-Heinz, Sanktionsverlangen im Wandel. Über das veränderte Verhältnis zwischen Schulbildung und Einstellung zur Todesstrafe (forthcoming).

<sup>24</sup> See Samuel A. Kirkpatrick for a study choosing attitudinal correlations as a dependent variable: Kirkpatrick, Age and Ideology: a Cohort Analysis of Attitude Constraint in the American Public, paper presented at the 1974 Annual meeting of the American Political Science Association, Chicago, August 29 – September 2 (1974).

### 2.3 Availability of Data

Cohort studies rest on the availability of comparable trend studies. Comparable trend studies might be originated and followed up in the future. Usually such a procedure does not provide immediate results and gratifications: If we want to follow up the life span and come to conclusions about aging effects in general and historically generated cohort effects on a comparative basis, it takes years and decades to obtain the results. Except in peculiar cases where periods of life only or shorter periods of time are of interest, for instance with regard to youth or times characterized by social movements, shorter spans of observation might be sufficient and even required. In the usual case, where longer time periods are relevant, we have to rely either on a secondary analysis of older surveys alone or on a combination of secondary analyses and primary analyses of surveys which represent *replications* of older surveys. Older surveys still exist to a large extent. The data sets have either been kept in the institutes themselves or have been given to a data archive. Especially in the United States there is still a rich data source going back as far as the 1930s, into the times of economic depression. Surveys spanning a broad time span and containing several identical or nearly identical questions have been conducted by a number of institutes, there, such as the Survey Research Center at the University of Michigan, the National Opinion Research Center (NORC) at the University of Chicago and the Gallup poll <sup>25</sup>. Of special interest seem to be the Gallup poll surveys in comparison to the other two, and since identical or nearly identical questions have been repeatedly asked, sometimes within a few weeks, a cumulation of surveys can be performed in order to have larger numbers of cases for subgroup analyses. The other two institutes have repeatedly asked only relatively few questions at intervals of several years. In case of the NORC data, moreover, the age has been coded in broad intervals until recently and has thus made cohort comparisons difficult.

Although the Gallup polls are the ones which provide a very rich data source, they are not without drawbacks: Like many of the samples drawn in the 1930s and 1940s the early Gallup polls were deliberately designed to underrepresent certain groupings of the population. „Since the earlier Gallup polls were used primarily to predict the outcome of elections and to provide readings on the sentiment of the electorate between the elections, the samples were designed to represent each population segment in proportion to the votes usually cast in elections rather than in proportion to the number of individuals in it. Since voter turnout was relatively low among females, Southerners, and persons with little education those segments of the population were deliberately underrepresented, and there were no southern blacks in many of the earlier samples.“<sup>26</sup> The change in representation took place gradually

<sup>25</sup> The following discussion of the U. S. surveys is based on Glenn, *Cohort Analysis*, pp. 34 passim.

<sup>26</sup> Op. cit., pp. 34 passim.

in the 1940's and 1950's. The first Gallup omnibus survey to use a sample with proportional sampling of females was in 1944 and the first omnibus survey with proportional representation of Southerners was not earlier than 1953. These changes in sampling procedures must have had effects on results and must therefore be taken into consideration, by weighting, for instance, when comparisons are made. In the 1950's Gallup, like other major survey organizations, switched from quota to random sampling.

In West Germany old data exist as well, though to a much smaller extent. The earliest non-local surveys were done right after the war on behalf of the American military commission by an institute which later became the DIVO institute. In the second half of the 1940's other survey institutes such as the Institut für Demoskopie and EMNID began their work as well. With the exception of the DIVO institute, which was abolished in 1969, the other institutes have continued their work until today. They have repeatedly asked a number of questions and thus have made trend studies possible. The original data sets, however, have only partially survived.

Concerning the survival of the old data the situation in Germany is not the very best. No data from the military government surveys still exist<sup>27</sup>. With regard to the DIVO institute a few surveys have been collected in the Zentralarchiv für empirische Sozialforschung in Cologne. The same applies to EMNID surveys. The Institut für Demoskopie remains to keep the richest data source with regard to old surveys. In contrast to the other commercial opinion and marketing research institutes it kept most of the original data sets from the beginning. Unfortunately it is not always easy to obtain these data for secondary analysis. The chance of performing trend and cohort analyses on the basis of older survey data alone thus does not look very promising. Under this condition the main road towards the study of trends and cohorts would be to take the old, archived data as a starting point and to conduct new surveys which are partially organized as replications of the older ones.

The preconditions for such strategies are quite satisfactory. The data quality of the older surveys in general does not seem to be too bad. There has been no deliberate attempt to have certain groups excluded or underrepresented as in the early Gallup surveys. The samples of the Institut für Demoskopie and the EMNID institute are usually quota samples which use sex, age and profession as quota criteria after the local sampling points have been determined. This has been so from the very beginning. Later developments have led to internal refinements, for instance with regard to combinations of variables like age and sex. The DIVO institute used random samples.

Regionally all states of West Germany have proportionally been included in the surveys of the diverse institutes from the beginning. Since the Saar was not part of West Germany until 1957 it was not represented in the early surveys. The population of the Saar constitutes a rather small proportion of the West German popula-

<sup>27</sup> See Allerbeck, Klaus R., *Demokratisierung und sozialer Wandel in der Bundesrepublik*, Opladen 1976, pp. 11 *passim*.



tion, however, and its later inclusion does not seem to have had any significant effect on the results.

With regard to the age categorization, which proves to be so important for the construction of cohorts, the situation in the surveys is worse. The Institut für Demoskopie used rather broad categories in the beginning. In 1950, for instance, only four age categories were employed (-30, 31-49, 50-65, 65+). Later, refinements were made. Today the age categorization is according to a five year interval. A similar broad categorization of age used to be common among the other institutes like EMNID and DIVO. EMNID has meanwhile opted for numerical coding.

## *2.4 Problems of Comparability*

Researchers successful enough to locate surveys with questions on comparable topics face a number of problems before they can actually start their research. They have to find out whether true comparability with regard to data collection and measurement techniques exists.

Comparability with regard to data collection might be limited. Quota samples, for instance, differ somewhat from random samples in their composition of respondents: Socially active respondents are usually more prevalent in quota than in random samples<sup>28</sup>. Even demographic comparability thus does not guarantee comparability beyond demographic variables. Consequently, caution has to dominate the interpretation when patterns of social activity (or variables associated with social activity) are compared in quota and random samples. Knowing the bias might, however, help in taking it into consideration. If a replication of older surveys is done, it might be wise to choose the same kind of sampling procedure: Using the same kind of sampling makes one feel sure that sampling is unlikely to be the cause of the observed differences.

Differences in data collection procedures might also derive from the interviewers. Black interviewers were not employed in American surveys at least through the early 1960's. Some institutes have meanwhile changed their policy and try to achieve racial matching in interviews nowadays<sup>29</sup>. Since race of interviewer might have effects on the respondents' answers when racially sensitive topics are concerned, the introduction of black interviewers might have caused an artificial change of opinion in surveys among blacks. To make the matter worse, racial interviewer effects could even change when racial matching is not altered: Since respondent-interviewer inter-

<sup>28</sup> v. Koolwijk, Jürgen, *Das Quotenverfahren: Paradigma sozialwissenschaftlicher Auswahlpraxis*, in: v. Koolwijk, Jürgen, and Wieken-Mayer, Maria (eds.), *Techniken der empirischen Sozialforschung*, Vol. 6, München and Wien 1974, p. 92.

<sup>29</sup> Glenn, *Cohort Analysis*, p. 32.

actions are colored by general interaction patterns among the races, changes in the relationship might affect respondent-interviewer interactions as well<sup>30</sup>.

The biggest problem with comparability of surveys, however, concerns question formulation and coding. Changes in question formulations might basically alter the distribution of responses and relationships between variables as well. Implicit differentials in the references of time as mentioned in the question wordings, might also influence the relevance of age, cohort or period effect on a given dependent variable. Compare for instance the subtle difference in the party identification question used by Gallup and the Survey Research Center: „In politics, as of today, do you consider yourself a Republican, Democrat or Independent” vs. „Generally speaking, do you usually think of yourself as a Republican, Democrat, Independent or what?”<sup>31</sup> Though there is something like functional equivalence of question formulations, the problem of identifying functional equivalence has not yet been solved. The proposal to take equivalent marginal<sup>32</sup> distributions at a given point in time as a criterion has been ruled out as invalid in the meantime: Even if marginal distributions are equivalent, relationships between variables might differ<sup>33</sup>. Caution also has to be exercised when major changes of potential relevance occur in the prior context of the question. Response categories, finally, seem to be of importance as well. Changes, however slight, could cause a shift to other categories<sup>34</sup>. It seems wise, then, to replicate questions and response categories exactly as they used to be, even if the formulations and categories do not seem to be ideal to the researcher. Otherwise it might be impossible to determine whether the change (or stability) in the responses is due to a real change or caused by variations in wording.

Open questions lend themselves to an especially large potential bias. Thus, a dramatic drop in antisemitic answers to an open question was found in the United States, for instance, when the surveying task was transferred from one polling agency to another<sup>35</sup>. Differences in answers to open questions were also found in one of the recent Detroit Area Studies (DAS) when compared with older ones. When the old answers to the questions were recoded by the staff who did the coding of the new surveys, the answers proved to be nearly identical in both points

<sup>30</sup> Schuman, Howard J., and Converse, Jean M., *The Effects of Black and White Interviewers on Black Responses in 1968*, in: *Public Opinion Quarterly*, 35 (1971), pp. 44–68.

<sup>31</sup> See Converse, *Party Support*, p. 35.

<sup>32</sup> Klingemann, Hand D., and Pappi, F. U., *Möglichkeiten und Probleme bei der Kumulation von Umfragen*: in: Wildenmann, Rudolf (ed.), *Sozialwissenschaftliches Jahrbuch für Politik*, Vol. 1, München and Wien 1969, pp. 173–190.

<sup>33</sup> Presser, Stanley, and Schumann, Howard, *Question Wordings as an Independent Variable in Survey Research: A First Report*. Paper presented at the American Statistical Associations meeting, Atlanta (August 1975), p. 5.

<sup>34</sup> Hyman, Herbert H., *Secondary Analysis of Sample Surveys*, New York 1972, p. 218; Glenn, *Cohort Analysis*, p. 29; Presser, *Question*.

<sup>35</sup> See Hyman, *Secondary Analysis*, p. 217.

of time<sup>36</sup>. The reasons for such divergences might be due to an insufficiently detailed code plan and a too heavy reliance on training and convergences during the prestages of coding or even in the stage of coding. The usual indicators of reliability — intercoder agreement — all too often reflect consensus after verbal socialization rather than consensus by strict application on an elaborated coding scheme itself. Unless exact procedures are found to assure high coding reliability, it might be wise to keep relevant old questionnaires on microfilm so as to allow later recoding.

### 3. Prospective Longitudinal Studies

#### 3.1 Characteristics

Cohort studies, as described above, are a fruitful instrument for tracing changes on an aggregate level over time. What they do not allow, however, are conclusions about change and stability on an individual level: Cohorts remaining constant in the distribution of attitudinal and behavioral patterns need not remain stable on an individual level as well (Table 3). In fact, it has quite often been shown that stability on an aggregate level is compatible with change on an individual level. And this individual change could be quite dramatic, in the extreme it could imply a complete turnover<sup>37</sup>.

In view of the possibility that aggregate change or stability does not say very much about the extent of change on an individual level, prospective longitudinal studies on an individual level are highly desirable. They not only allow measuring the extent of change, but also tracing individuals' attitudes and behavior patterns over time and establishing causal relationships between antecedent conditions and subsequent outcomes. Problems of selective exposure and recruitment or recursive influences, such as in cross-sectional studies, are easier to deal with<sup>38</sup>.

<sup>36</sup> Duncan, Otis D., et al., *Social Change in a Metropolitan Community*, New York 1973, pp. 21 passim.

<sup>37</sup> For individual turnover and a stability of marginals see for example Converse, Philip E., *Attitudes and Nonattitudes, Continuation of a Dialogue*, in: Tufte, E. R. (ed.), *The Quantitative Analysis of Social Problems*, Reading/Mass. 1970, p. 170; Butler, *Political Change*, p. 221.

<sup>38</sup> This does not imply that longitudinal studies are always advantageous to cross-sectional studies. For a more general discussion see Dierkes, Meinolf, *Die Analyse von Zeitreihen und Longitudinalstudien*, in: v. Koolwijk, Jürgen, and Wieken-Mayser, Maria (eds.), *Techniken der empirischen Sozialforschung*, Vol. 7, München and Wien 1977, pp. 133 passim; Goldfarb, *Introduction*.

Table 3: Aggregate and Individual Level Stability:Hypothetical Examples (Absolute Numbers)

### Stability on the Aggregate Level

		t <sub>1</sub>		Total
		A	B	
t <sub>2</sub>	A	40	—	40
	B	—	60	60
Total		40	60	100

$Q = +1.0$

		t <sub>1</sub>		Total
		A	B	
t <sub>2</sub>	A	—	40	40
	B	40	20	60
Total		40	60	100

$Q = -1.0$

### Instability on the Aggregate Level

		t <sub>1</sub>		Total
		A	B	
t <sub>2</sub>	A	40	20	60
	B	—	40	40
Total		40	60	100

$Q = +1.0$

		t <sub>1</sub>		Total
		A	B	
t <sub>2</sub>	A	—	60	60
	B	40	—	40
Total		40	60	100

$Q = -1.0$

Prospective longitudinal studies not only have positive methodological aspects, however. They lack a control of cohort and period effects:

1. Longitudinal studies covering a broad time span often rely on one cohort only — sometimes consisting of individuals born within a certain week of a particular year. The problem here is, that we do not know whether other cohorts really undergo the same development. It might well be that individuals who have experienced certain generational imprints, vary from others in their specific ways of life. On cross-sectional studies the problem is not better. Though we now have several cohorts and can follow them up as they age, we do not have any further incoming cohorts, as in repeated cross-sectional sampling for comparisons.

2. The other problem concerns period effects. Without additional information, it is often impossible to separate aging and period effects. Increases of income with advancing age over time might, for example, have been caused by processes associated with the life cycle or they might have been affected by period trends which in-

creased the general income levels in the population. Probably many studies have not paid adequate attention to these period effects and have therefore come to questionable conclusions.

The only possibility of controlling for period effects in a longitudinal design would be to combine it with a repeated cross-sectional design. At first sight this kind of approach might appear to be rather costly. However, if the number of dependent variables deemed important is sufficiently small, the use of omnibus surveys for locating a few questions in trend studies would be sufficient and could keep the costs down. Under certain conditions the problem of assessing period influences could also be solved by taking questions which are repeatedly asked by the big survey organizations as a starting point. Incorporating these questions into one's questionnaire would allow a comparison with the periodically published findings of the survey organizations.

### *3.2 Possible Designs and Their Problems*

Researchers interested in launching a prospective longitudinal study face a number of options and problems<sup>39</sup>. They might decide to begin an entirely new series of studies or they might decide to follow up a study that has already been begun by other researchers in the past.

Researchers deciding for a new prospective longitudinal study might need a long time perspective. When whole life histories are the subject, it takes years until all the desired data have been collected. The researcher is likely to have lost interest or might have even died before the study's original aims are accomplished. Similar problems might even exist in cases of projects concentrating on certain phases of the life cycle. For this reason the project organization must be such as to safeguard continuity beyond the individual researcher. Depending on the time span, hypotheses and interests might also change during the course of research. It has therefore been suggested that one have a rather broad approach in the beginning and collect as much information as possible in order to have more options for research inquiries at a later date. Such a strategy, however, has its dangers. It might have a too global approach as the result and lack an adequate theoretical concept. Needed is the delicate balance between broadness and selectivity.

Whether the use of local or national samples is preferable, has to be decided individually. Local studies very often offer the „advantage of greater homogeneity in data collection, the possibility of utilizing locally available specialized resources for more thorough examination, and of augmenting samples of the comparatively rare

<sup>39</sup> See the discussion in Wall, *Longitudinal Studies*.

categories"<sup>40</sup>. More administrative data are at hand and more easily accessible than national samples would allow.

A different approach to conducting longitudinal studies is to proceed from data which have already been collected. In a number of cases — especially where panel designs were used — addresses might still exist. One could start from them and re-visit the project by following the individuals who have once been researched. In other cases rather old longitudinal studies encompassing long time spans might exist. Depending on the interest into socio-historic circumstances and their effects, the old data might either be reanalyzed or combined with new, freshly collected data.

Older longitudinal studies available for such purpose are probably more numerous than is usually supposed. In the 1920's and 1930's several longitudinal designs were begun, especially in the United States and mostly with regard to child training and development<sup>41</sup>. Others, starting at about the same time, focused on medical public health or — as a number of Swedish ones — on the interplay of social background, intelligence and schooling for one's later career<sup>42</sup>. A few more recent studies have focused on the process of coming into old age or on topics like delinquency. In the latter case administrative data and, here and there, a combination of administrative and interview data have been used<sup>43</sup>.

Perhaps the most interesting studies are those which have been done in the child training and personality development field. They seem of great interest since many of them go very far back in time and were still in use not too long ago. Their prospects for research seem great: They offer the unique possibility of showing how life chances depend on historical circumstances and on one's location in the social structure. At first sight, however, the merits for social and socio-historic research might seem limited. The studies are based „upon small numbers of subjects chosen on the more or less fortuitous basis of contiguity to the research center and the willingness to cooperate“<sup>44</sup>. Emphasis is placed on a biological model of child growth, on somatic growth and measures of personality attributes. Attitudinal and behaviour patterns receive less attention and socio-cultural and environmental influences are hardly measured<sup>45</sup>. Information about the parents is only partially

<sup>40</sup> Wall, op. cit., p. 67.

<sup>41</sup> For a list of the early longitudinal studies see Wall, op. cit.

<sup>42</sup> See Janson, Metropolitan; Goldfarb, Introduction.

<sup>43</sup> E. g. Wolfgang, Marvin E., et. al., *Delinquency in a Birth Cohort*, Chicago 1972; Jacoby, Joseph E., et al., *Drug Use and Criminality in a Birth Cohort*, in: *Drug Use in America. Problem in Perspective*. Technical papers of the second report of the National Commission on Marihuana and Drug Abuse, Vol. 1, Washington 1973, pp. 300–372; West, Donald J., and Farrington, David P., *Who Becomes Delinquent*, London 1973.

<sup>44</sup> Wall, *Longitudinal Studies*, p. 32.

<sup>45</sup> See Neugarten, Bernice L., Introduction to the Symposium: Models and Methods for the Study of the Life Cycle, in: *Human Development*, 14 (1971), pp. 84 passim; Bloom, Benjamin S., *Stabilität und Veränderung menschlicher Merkmale*, Weinheim 1971, pp. 14, 201 passim (originally published in English as *Stability and Change in Human Characteristics*, New York 1964).

sought. If the parents were interviewed themselves, it used to be the mother — she was apparently seen as the most important figure in the child's world.

Despite these individualistic, heavily psychologically and biologically oriented approaches, however, many studies also contain some useful and substantial information on family life, school and social experience and sometimes even on the larger milieu at that time. Performing a secondary analysis, or — under certain circumstances — a secondary *and* primary analysis of new data could provide new insights. And this insight could be very meaningful to a socio-cultural and historical approach as Glen Elder has shown in a number of brilliant analyses<sup>46</sup>. Hereby special efforts have, of course, to be placed in finding appropriate indicators for the measurement of the impact of historical events: One has to look for the available variables and make the best of them. Elder for instance chooses loss of family income during the depression years as an indicator, which links historical circumstances — economic depression — to the individual's life course: People who recorded a loss of income were treated as being affected by the economic depression as a historical event. Whether the use of this indicator alone is sufficient might be arguable<sup>47</sup>, but this has to be seen against the background of available indicators in the study.

### *3.3 The Problem of Respondent Loss*

Where longitudinal studies are begun or continued the most serious practical problem usually to be solved is retrieving the respondents. Depending on the characteristics of the population, the records with which one is starting, and the time that has elapsed, the number of cases which appear to be lost at the outset is normally quite high. Many researchers have therefore seen longitudinal studies as more or less fruitless endeavours from the very beginning. „However, most of these cases *can* be retrieved“, a review indicates, and „if longitudinal study comes out badly, it usually is because little or no effect was made“<sup>48</sup>. Numerous examples can be quoted where more than 90 % of the subjects were retained. And these

<sup>46</sup> Elder, Glen H., *Children of the Great Depression*, Chicago 1974; Elder, Glen H., and Rockwell, Richard, C., *Economic Depression and Postwar Opportunity in Men's Lives: A Study of Life Patterns and Health*, in: Simmons, Roberta G. (ed.), *Research on Community and Mental Health. An annual compilation of research*, Greenwich/Conn. 1977.

<sup>47</sup> See Fürstenberg, Jr., Frank F., *Review Essay*, in: *American Journal of Sociology*, 81 (1975), p. 651.

<sup>48</sup> Eckland, Bruce D., *Retrieving Mobile Cases in Surveys*, in: *Public Opinion Quarterly*, 32 (1968), p. 53.

studies having such great success rates prove to be by no means unique studies with regard to population samples and design strategy. Even with representative population samples based on random sampling high success rates have been achieved over the years<sup>49</sup>.

There can be no doubt, nonetheless, that some studies — due to sample and design characteristics — have greater chances than others to succeed, regardless of the techniques used to locate respondents. Major factors affecting tracing failures are the size, mobility and dispersion of the sample. Among social groups relatively homogeneous and centrally located populations with high education (such as university students) the researcher probably fairs best. The chance of locating them might be high simply because people tend to stay in these institutions for some time, establish contacts with many people there, and later follow certain relatively homogeneous career patterns. They might, moreover, belong to alumni organization and to professional organizations, whose directories might be searched. High education might furthermore increase cooperativeness.

Of further importance seems to be the elapsed time between contacts either with the researcher himself or some organization whose records are checked by him: Contacting people within relatively short time intervals reduces the chances of losing geographically mobile persons out of sight. In cases of a contact between researcher and respondent the researcher might learn about moves in advance or might at least still obtain information from either the neighborhood or public administration (such as postal service) just in time to follow up the paths of mobile respondents. In cases of deviants (such as criminals or drug users) those heavily engaged in the deviant pattern will have a greater chance of becoming officially visible to one of the organizations dealing with them (such as police, treatment centers, prisons). They will leave more traces in the official records to which the researcher has access<sup>50</sup>.

<sup>49</sup> Eckland, *op. cit.*, pp. 53 *passim*; Parnes, Herbert S., *Longitudinal Surveys: Prospects and Problems*, in: *Monthly Labor Review*, 95 (1972), pp. 14 *passim*; Wall, *Longitudinal Studies*.

<sup>50</sup> For a discussion of selective respondent losses in cases of deviance research — especially on drug use — see Nurco, David, et al., *Locating Respondents*, in: Johnston, Lloyd D., et al. (eds.), *Conducting Follow up Research on Drug Treatment Programs*, NIDA Treatment Program Monograph Series No. 2, Washington 1977, pp. 72 *passim*.



### 3.4 Minimizing Attrition Rates

Systematic reviews of the techniques and practical approaches to lowering attrition rates by choosing appropriate searching procedures have, as yet, been scarce<sup>51</sup>. Without doubt, the best way to approach this problem is to think of it in the beginning of the project and to incorporate appropriate devices. Whether the use of volunteers constitutes a solution to the problem — as has been suggested by some writers — however, remains doubtful: Volunteers usually do not constitute a representative sample. They deviate from others in social, behavioral and attitudinal characteristics<sup>52</sup>. Random samples in general seem better suited for the purpose. In this case the most useful approach is to collect extensive biographical information (which could provide cues for later searching procedures) in the first interview. It could and actually should even be done when a longitudinal design is not planned: Doing so opens up the possibility of reinterviews one day, and having extensive biographical information makes the chances of tracking down respondents much better<sup>53</sup>. Where public records have to be checked in the course of the study — either to locate respondents or to assess certain career patterns — written permission to use these records might be useful: Privacy legislation has made it much more difficult for researchers in many countries to obtain information from available records. In cases where the evaluation of treatment success is of central importance for organizations (such as treatment centers) it has therefore been suggested that people be asked to sign such permissions when entering the institution. Other strategies to lower attrition rates could be to ask the respondents to notify the researcher of any change of address. Respondents might be motivated for such cooperation by keeping them informed about the progress of the project. In a Swiss project, for example, parents whose children were followed up in a longitudinal study are periodically informed by newsletters about some results of the surveys<sup>54</sup>. This kind of feedback might have the effect of increasing the attachment to the project and the degree of cooperation extending to both information about migration and willingness to be interviewed.

<sup>51</sup> For some exceptions see Clarridge, Brian R., et al., *Tracing Members of a Panel: A 17 Years Follow-up*, Center for Demography and Ecology, University of Wisconsin, CDE Working Paper, No.72-73; McAllister, Ronald J., et al., *Tracking Respondents in Longitudinal Surveys: Some Preliminary Considerations*, in: *Public Opinion Quarterly*, 37 (1973), pp. 413-416; Maddux, James F., and Desmond, David P., *Obtaining Life history Material about Opioid Users*, in: *American Journal of Drug and Alcohol Abuse*, 1 (1974), pp. 181-198; Eckland, Retrieving; Nurco, Locating respondents.

<sup>52</sup> See for instance Riegel, Klaus F., et al., *The Prediction of Retest Resisters in Research on Aging*, in: *Journal of Gerontology*, 23 (1968), pp. 370-374.

<sup>53</sup> McAllister, Ronald J., et. al., *Tracking Respondents*; Eckland, Retrieving; Nurco, Locating Respondents.

<sup>54</sup> Project directed by Prof. Ries, now at the University of Trier, W. Germany.

Especially in this kind of approach, however, the researcher also has to be extremely careful of not exerting any influence which might cause artificial results.

Whatever strategies are chosen to minimize loss of respondents over time, techniques to locate sample losses have to be employed in most of the studies. Some possibilities exist and have proved fruitful. They will, of course, vary between countries depending on registration systems and the distribution of certain characteristics in the population (such as telephone ownerships). Postal services, including telephone services, public records and private services, including organizational directories, could be used. Public records seem to be the ideal first source to be employed because they are so readily available. One might begin by using the city directories and the telephone directories plus the help of public administrations (such as the Meldeämter in Germany) or telephone information operators to find the subjects or relatives. If this fails one might — where available — also use street directories to identify former neighbors. Other sources of help might be post offices, where records of address change may be kept for some time, and in case of deviants, social control agencies with which the individual might have come into contact meanwhile. Finally, neighbors and friends may know about the new address, relatives — especially parents — will certainly do so. The search for new addresses might even be included into the research designs itself. Thus, in a British study of former students, a list of the people graduating with the respondents at the same university was included with the questionnaire and people were asked to fill in current addresses of those with whom they were still in contact<sup>55</sup>.

## 4. Retrospective Studies

### *4.1 Characteristics*

Retrospective interviewing is often the only possible means to secure life history material on an individual level. If longitudinal studies do not exist, no other method is available. In many cases, moreover, longitudinal studies cannot be performed since the respective groups of interest occur too seldom in the population. Under these circumstances, the inauguration of a longitudinal study on the basis of a representative sample would yield too few cases. But retrospective interviewing is not only an alternative to prospective longitudinal studies. Retrospective interviewing is

<sup>55</sup> Chishom, Lynne, et al., *Methodological Problems of Quantitative and Qualitative Research: the National Survey of British Graduates*, in: *Angewandte Sozialforschung*, 5 (1977), pp. 196 passim.

usually a fruitful complement to panel and other prospective longitudinal studies, since these studies only provide measurements at selected points in time. The processes that have happened externally and internally to the actor in the meantime cannot be measured except by retrospective assessment.

Retrospective assessment might even be more than a method which fills the gaps other methods leave. Retrospective assessment, if properly done, offers the unique opportunity to unravel the complex process by which patterns of attitudes and behavior are elaborated and transformed in response to external events. It allows the analysis of events impinging on an individual's life course and the definitions which are directed to them, and hence the analysis of interplay of objective and subjective factors. It is the only method „if one wants to know how an action came to be — what the key choices were; what the actor thought he was doing and how he felt about it; what influences were present and what triggered the action; and what outcome the actor expected“<sup>56</sup>.

Finally, retrospective interviewing has been a tool used to collect information about life in earlier historical periods. Interviewing old people has been valued for the richness and vividness of details relating to past days and for its extensive coverage of events and experiences. The respondents have been seen as the carriers of memories of the past — as experts, who can provide material which cannot be secured by ordinary historical approaches. By making people remember their own personal life history and, additionally, the general circumstances at different points in time, attempts are made to reconstruct the past from „oral history“<sup>57</sup>.

#### *4.2 Early Approaches*

Interest in life histories on the basis of retrospective interviewing is nothing new to the literature. It has quite a long, though interrupted, tradition especially in psychology and sociology. Life history approaches flourished in the 1920's and 1930's in Europe and the United States. In Europe it was Charlotte Bühler who took a major interest in life histories from a psychological perspective. For her, the understanding of a human being as a whole implied the knowledge of his whole life history. Major emphases were placed on a combination of developmental and motivational approaches, on life stages and sequences, on the realization of aims and the inter-

<sup>56</sup> Kadushin, Charles, Reason Analysis, in: International Encyclopedia of the Social Sciences, Vol. 13, New York 1968, p. 338.

<sup>57</sup> See for instance Thompson, Paul, The Voice of the Past. Oral history. Oxford 1978; Barker, Diana, and Thompson, Paul, The Comparative Study of Courtship and Marriage in Twentieth Century Britain: A Methodological Discussion. Paper presented at the 13th Seminar of the Committee of Family Research of the International Sociological Association, Paris 1973.

play of objective and subjective factors<sup>58</sup>. At about the same time in the United States, William Thomas took an interest into life histories from a sociological perspective. For him, life histories based on biographical material constituted the ideal approach to understanding the interplay between objective and subjective factors, and the sequential development of attitudes, behavior patterns and personality traits over time. For him also, any specific act becomes comprehensible only in the light of its relation to the sequence of past events<sup>59</sup>.

Unfortunately, Bühler and Thomas, did not succeed in elaborating their life history approach. Bühler remained conceptually and methodologically in an early stage. The impact of her work on psychology seems to have been minor. Life history studies there are still scarce<sup>60</sup>. Thomas in general provided a better approach systematically and methodologically. However, he failed to integrate his life history approach adequately with his empirical material, and did not go beyond a mere presentation of individual life histories. Quantitative analyses which allow a more analytical approach were not employed. His failure to make his methodology more rigorous and his over-reliance on written autobiographies rather than specially designed interviews probably doomed his approach to failure<sup>61</sup>. It has only been in Poland where his collaborator Znaniecki gave the biographical approach a major impetus. Nonetheless, breakthroughs seem to have been rare<sup>62</sup>. In the United States, his approach finally died out with the influence of the Chicago school of sociology.

Apart from the Bühler and Thomas tradition centering on life histories as a whole, there have been only a few other attempts towards a life history approach. In these cases, the approach was more limited, aiming at an understanding of decisions and selected career phases within the life history. Paul F. Lazarsfeld, a former collaborator of Charlotte Bühler, has been one of the most important contributors. In a paper published in 1934 he gives an outline of his approach to understanding marketing decisions. He does so in a way which contains many of the central ideas of Thomas, without probably knowing of them. „The action“, he writes, „is a joint product of factors in the individual and factors in the situation. Explanations must always include both the objective and the subjective, and these are always inseparable interrelationships“. Furthermore: „At each successive stage

<sup>58</sup> See Bühler, Charlotte, *Der Lebenslauf als psychologisches Problem*, 2nd rev. ed., Göttingen 1959. For later statements on the approach see Bühler, Charlotte, and Massarik, Fred (eds.), *Lebenslauf und Lebensziele*, Stuttgart 1969.

<sup>59</sup> Thomas, William I., *Person und Sozialverhalten* (ed. by Edmund H. Volkart), Neuwied 1965 (originally published as *Social Behavior and Personality: Contributions of W. I. Thomas to Theory and Social Research*, New York 1951).

<sup>60</sup> Bühler, Charlotte, *Basic Theoretical Concepts of Humanistic Psychology*, in: *The American Psychologist*, 26 (1971), p. 380.

<sup>61</sup> For a critique see Madge, John, *The Origins of Scientific Sociology*, New York 1962, pp. 52 passim; Fürstenberg, Review Essay, p. 648.

<sup>62</sup> See Bukowski, Jacek, *Biographical Method in Polish Sociology*, in: *Zeitschrift für Soziologie*, 3 (1974), pp. 18–30; Szczepanski, Jan, *Die biographische Methode*, in: König, Rene (ed.), *Handbuch der empirischen Sozialforschung*, Vol. 4, 3rd ed., Stuttgart 1974, pp. 226–252.

along the time-line . . . we are dealing with a changed person, different by reasons of what has occurred at preceding stages and, also, of course, we have new influences that have come into operation to affect the on-going action"<sup>63</sup>. Lazarsfeld's conceptualization of what has been termed „reason analysis“ has had some influence on research, mostly among his students at Columbia. In the long run, his approach, however, for whatever reasons, has remained without major impact<sup>64</sup>.

Besides Lazarsfeld there have been approaches to the study of careers in the deviancy field. The most prominent is Howard S. Becker, who, in 1963, stressed the need to have sequential rather than simultaneous models of deviance. Patterns of behaviour develop in orderly sequence, he writes, and we must deal with the sequence of steps, of changes in the individual's behaviour and perspectives, in order to understand the phenomenon. „Each step requires explanation, and what may operate as a cause at one step in the sequence might be of negligible importance at another step“<sup>65</sup>. In developing his approach, which remained relatively rudimentary, Becker might have been influenced by Thomas and his followers working on life histories. We cannot say with certainty<sup>66</sup>. Anyhow, Becker has probably been the only one of those mentioned, who still has some impact on research today — though in a limited way: His career approach is mostly applied to that phase in the delinquent's career which starts *after* law enforcement has got hold of him. Earlier phases are hardly dealt with under a similar perspective.

As a whole then, the abovementioned approaches, centering on life histories or stages in life histories, have not had far reaching influence. They have fallen into disuse to a large extent or have been rather selectively used (as in Becker's case). Systematic further developments have not taken place and, above all, methodological improvements have hardly been made. „Considerable work remains to be done“, writes for instance Charles Kadushin in reviewing the Lazarsfeld approach, „in enhancing and testing the validity of reason analysis, as well as in testing and improving various models of action“<sup>67</sup>. Hence more than 30 years after its development,

<sup>63</sup> Kornhauser, Arthur, and Lazarsfeld, Paul F., *The Analysis of Consumer Action*, in: Lazarsfeld, P. F., and Rosenberg, M. (eds.), *The Language of Social Research*, Glencoe/Ill. 1955, pp. 393, 396. See also Lazarsfeld, P. F., *The Art of Asking Why*, in: Katz, D., et al., (eds.), *Public Opinion and Propaganda*, New York 1954, pp. 765—786.

<sup>64</sup> See also Lazarsfeld, Paul F., *Eine Episode der empirischen Sozialforschung*, in: Parsons, Talcott, et al., *Soziologie — autobiographisch*, Stuttgart 1968, p. 215. The lack of long ranging impact is remarkable. Though some excellent research projects have been based on „reason analysis“ (for instance Rossi, P. H., *Why Families Move*, 1955) no further developments have taken place. Neither Lazarsfeld himself nor his students seem to have had much interest in it at a later date.

<sup>65</sup> Becker, Howard S., *Outsiders*, New York 1973, originally 1963, p. 23.

<sup>66</sup> Becker does not mention Thomas, but he must have been aware of his approach. See for instance his foreword to C. Shaws' book *The Jack Roller*, reprinted in Becker, Howard S., *Sociological Works. Method and Substance*, Chicago 1970, pp. 68 *passim*.

<sup>67</sup> Kadushin, *Reason Analysis*, p. 342.

reason analysis has still not proceeded very far. The same is even more true of the other approaches. Retrospective life history studies have failed to develop a tradition of their own.

### *4.3 Problems of Recall*

The usefulness of retrospective studies for reconstructing life histories largely rests on the validity of the data which might be obtained. However, the validity of recall has only been given scant attention in past approaches to the study of life histories. For W. Thomas and followers „even the highly subjective record has a value for behaviour study . . . for (the subject's) immediate behaviour is closely related to his definition of the situation, which may be in terms of objective reality or in terms of a subjective appreciation“<sup>68</sup>. This might, without doubt, be true for present behavior. But if one intends to go back and to reconstruct the whole life history, such an attitude does not make much sense. Recollections of past subjective definitions of situations might undergo a distortion by memory. For this reason an adequate reconstruction of the life history without considerations of methodological factors must remain problematic. In the other approaches to the study of life histories, apart from the Thomas tradition, problems of recall usually are not even mentioned.

In view of the lack of methodological research directed to the retrospective study of life histories, we have to rely on the rather scattered and still rudimentary findings from the survey research literature dealing with problems of recall. The methodological studies on retrospective interviewing here usually do not pertain to sequences of events or life histories as a whole, but to the recollections of single experiences or events only.

The results of this research on the validity of retrospective interviewing are complex and, at first sight, rather perplexing. Whereas the validity of recall must be seen as high in some cases, it proves to be low in others. Closer looks at the studies reveal a number of factors that intervene and produce varying results. One seems to be the kind of subject: attitudes are probably more difficult to remember than practical matters, behaviour patterns or events<sup>69</sup> — especially so, if the attitudes are short-term attitudes and not very much an object of reflection. If one realizes how much attitudes fluctuate over time, it really is no wonder that people have difficulties with recall<sup>70</sup>. Recall of attitude data might be somewhat better when it

<sup>68</sup> William Thomas as quoted in Shaw, Clifford R., *The Case Study Technique. Value of Delinquent Boys own Story*, in: Wolfgang, Marvin E., et al. (eds.), *The Sociology of Crime and Delinquency*, New York 1962, p. 102.

<sup>69</sup> Barker, D., and Thompson, P., *Courtship*, p. 16.

<sup>70</sup> On the instability of attitudes see for example Converse, *Attitudes*.

is not the description of attitudes at certain dates which is required, but directions of attitude change. But even here the relationship between perceived and actual attitude change is usually low<sup>71</sup>. It might be only among certain segments of the population, as in highly educated educational groupings, that the direction of the change is adequately perceived. Greater ideological involvement could account for this pattern<sup>72</sup>.

Aside from kind of subject, factors associated with the subjective relevance of the topic, affect the validity of recall. Amount of elapsed time, past and present salience and past and present subjective threat have an impact. The validity of data decreases as time increases and saliency decreases. Experiences, which were prominent once are likely to be forgotten if they have little relevance to the respondent's current life. Not all serious events are, however, equally kept in mind. If experiences have been threatening (as in certain cases of surgery) the rate of recall tends to be lower. Recollections of extremely negative experiences are suppressed<sup>73</sup>. In which direction memory distortions finally go depends on various factors. One has to do with time location, others with normative and cultural expectations and culturally induced frames of reference, against which experiences are evaluated. If one takes continuous states, such as income or certain attitudes, and compares present and past recollected ones by controlling for past conditions, a distortion towards present states can be found, i. e. past income or attitudes are nearer to the present than to the past. The past is colored by the present<sup>74</sup>. With regard to single events (such as being in a hospital) a tendency to shift the reported item into or out of the recall period (telescoping effects) exists. In most cases the effects seem to be in the forward direction. This means, that events are mentioned as having happened in the recall period which actually happened before<sup>75</sup>.

Research also shows that a tendency exists to fulfil normative expectations by distorting actions which are negatively valued with one's reference group. Parents tend to report about their past child-rearing practices in accordance with what is normatively expected by the current child-rearing literature. Alcohol consumption

<sup>71</sup> Fink, T., The Retrospective Question, in: *Public Opinion Quarterly*, 24 (1960), pp. 143–148; Hardin, Elinar, Perceived and Actual Change in Job Satisfaction, in: *Journal of Applied Psychology*, 49 (1965), pp. 363–367.

<sup>72</sup> Newcomb, Theodore M., et al., *Persistence and Change: Bennington College and its Students after Twenty-Five Years*, New York 1967.

<sup>73</sup> Cannell, Charles, and Kahn, Robert L., Interviewing, in: Lindzey, G., and Aronson, E. (eds.), *The Handbook of Social Psychology*, Vol. 2, 2nd ed., Readings/Mass. 1968.

<sup>74</sup> Cannell, Interviewing.

<sup>75</sup> Neter, John, and Waksberg, Joseph, *Response Errors in Collection of Expenditures Data by Household Interviews: An Experimental Study*. Bureau of the Census, Technical Paper No. 11, Washington 1965, p. 12; Cash, Williams S., and Moss, Abigail J., *Optimum Recall Period for Reporting Persons Injured in Motor Vehicle accident*, U.S. Dept. of Health Education and Welfare, Vital and Health Statistics, Series 2, No. 50, Washington 1972, p. 8. No forward telescoping is found in Cannell, C. F., et al., *Reporting of Hospitalization in the Health Interview Survey*, Vital and Health Statistics, Series 2, No. 6, Washington 1965, p. 32.

is underreported and the visiting of union meetings is overreported by its members<sup>76</sup>. Whether there is a general tendency to shift negative experiences into the direction of more positive ones and to see the past more positively than it was, as has been suggested, is not quite clear.

Cultural expectations of „what goes with what“ affect recall as well. The recall of behavioral configurations, such as interaction patterns, might therefore often reflect primarily the subject's construction of the reality and not the world as it is<sup>77</sup>. For this reason the reconstruction of past events, especially if not highly associated with the subject himself – as in historical reconstructions by retrospective interviewing – might be problematic.

Finally historical factors could have an impact. They could change from the frame of reference against which experiences are evaluated. Germany serves as an example. One can observe there a continuous trend from indicating a happy childhood to indicating a partially happy childhood over the years. Breakdowns of the results show that this trend also holds for cohorts (Table 4). Apparently a general trend in the population has caused a reevaluation of childhood experiences. Shifting conceptions of what is the desirable child training, as observed in this period, might be the reason<sup>78</sup>.

*Table 4: Recollections of Childhood  
Percentage Indicating a „Happy Childhood“*

Age	Year of Survey		
	1956	1960	1974
16–29	63	58	67
30–44	63	58	57
45–59	57	56	44
60 and older	62	57	43
Total	61	57	53

Source: Weltz, F. (ed.), *Jugend im Sozialisierungsprozeß. Tabellenband*, München 1967, Table VI 1, VI 2 (based on surveys from the Institut für Demoskopie); Noelle-Neumann, E., *Allensbacher Jahrbuch der Demoskopie 1974–1976*, Wien and Zürich 1976, p. 8.

<sup>76</sup> Cannell, Interviewing, p. 545; Wall, W. D., and Williams, H. L., *Longitudinal Studies*, p. 9; Danziger, Kurt, *Socialization*, Harmondsworth 1971, p. 146; Dean, Lois R., *Interaction, Reported and Observed: The Case of One Local Union*, in: *Human Organization*, 17 (1958), pp. 35–44.

<sup>77</sup> D'Andrade, Roy G., *Cultural Constructions of Reality*, in: Nader, Laura, and Maretzki, Thomas (eds.), *Cultural Illness and Health*, Washington 1973, pp. 115–127; D'Andrade, Roy, *Memory and the Assessment of Behavior*, in: Blalock, W. (ed.), *Measurement in the Social Sciences*, Chicago 1974, pp. 159–186.

<sup>78</sup> See EMNID Informationen, 26 (1974), p. 7.



A still unsolved problem is whether recall patterns vary according to sociodemographic groupings. Very little information unfortunately exists in this field. And where results are available, they are mixed and contradictory. Some studies, for instance indicate worse recall patterns among lower classes, others do not. Some similarly show effects of age or sex and others do not show any such effects<sup>79</sup>. Probably social categories do not have any effects per se, but only when other recall relevant factors, such as subjective salience, are associated with them.

#### 4.4 *Optimizing Recall*

„,Remembered‘, ‚forgotten‘ and ‚never known‘ should not be regarded as absolute categories. Forgotten material can often be recalled with sufficient effort, and personal records are often available if respondents wish to consult them“<sup>80</sup>. Methodological strategies can be developed to cope with the problem more effectively.

One of these strategies concerns the formulation of questions and the building up of question sequences: The rate of recall usually increases as more specific and focused questions are posed and the better a match is accomplished between the categories of the questions and the manner in which recall of events is symbolically structured<sup>81</sup>. Stimulation and reinforcement by the interviewer could also help in producing a stronger motivation and more intense searching for recollections of past events. Mail and telephone surveys probably do not turn out equally well in this case because of the great importance of having an interviewer fulfil probing and reinforcing functions<sup>82</sup>. The more time the respondent has to remember, the more

<sup>79</sup> Brekstadt, Arne, Factors Influencing the Reliability of Anamnestic Recall, in: Child Development, 37 (1966), pp. 603–612; Hardin, Einar, and Hershey, Gerald L., Accuracy of Employee Reports on Changes in Pay, in: Journal of Applied Psychology, 44 (1960), pp. 169–175; Cukier Robins, Lilian, The Accuracy of Parental Recall of Aspects of Child Development and of Child Rearing Practices, in: Journal of Abnormal and Social Psychology, 66 (1963), pp. 261–270. For a discussion of assumed different function played by the memory in different age groups see Lieberman, M. A., and Falk, Jacqueline M., The Remembered Past as a Source of Data for Research on the Life Cycle, in: Human Development, 14 (1971), pp. 132–141; Barker, Courtship, p. 17.

<sup>80</sup> Cannell, Interviewing.

<sup>81</sup> Biderman, Albert D., The Distortions of Victimization Data and Mnemonic Effects, mimeo, Washington 1970.

<sup>82</sup> For interviewer activity see for example Marquis, Kent H., et al., Reporting Health Events in Household Interviews. Effects of Reinforcement, Question Length and Reinterviews; U.S. Dept. of Health, Education and Welfare, Vital and Health Statistics, Serie 2, No. 45. For the differential effects of personal interview, mail and postal interview see for example Turner, Anthony G., Methodological Issues in the Development of the National Crime Panel Survey: Partial Findings, mimeo., Washington 1972.

recollections are produced. If reinterviewing is done, moreover, often respondents who have been unable to remember past events, now report them. Many of the biases are hereby reduced<sup>83</sup>. A strategy to be explored — especially if interviewing is done with the same persons over a period of time (as in some interviews by historians) — would be to discuss again some of the events reported on an earlier occasion in order to activate more recollections. Bounded recall and use of anchoring points for the temporal reference period that also order experiences and the round of life („Since New Year's Day“) could be used to reduce telescoping effects<sup>84</sup>.

Finally, it has been found that asking individuals about the experiences of other members of their household (for example with regard to victimization or expenditures) produces less information than asking the other persons themselves. Asking the persons themselves and, under certain conditions, more than one person in the household, would therefore yield the best information: the others could help to recollect the events, experiences and attitudes<sup>85</sup>. Further strategies of obtaining better rates of recall exist and need further explorations. Though the situation with retrospective interviews is not the best in general, retrospective questions are often indispensable. More methodological attention should accordingly be directed to them.

Rearing Practices, in: *Journal of Abnormal and Social Psychology*, 66 (1963), pp. 261–270. For a discussion of assumed different functions played by the memory in different age groups see Liebermann, M. A., and Falk, Jacqueline M., *The Remembered Past as a Source of Data für Research on the Life Cycle*, in: *Human Development*, 14 (1971), pp. 132–141; Barker, Courtship, p. 17.

<sup>83</sup> Neter, Response errors, p. 3; Cannell, Hospitalization, p. 42.

<sup>84</sup> See for example Biderman, Albert D., *Surveys of Population Samples for Estimating Crime Incidence*, in: *The Annals of the American Academy of Political and Social Science*, Vol. 374, No. 21 (1967); Biderman, Distortion; Neter, Response Errors.

<sup>85</sup> Neter, op. cit., p. 3, 71; Biderman, Estimating; Turner, Methodological Issues.

The Analysis of Life Histories:  
Illustrations of the Use of Life History Plots\*

In the last few years analysis of life histories has been rediscovered as a sociological problem area from a variety of different perspectives. For a long time the classical study of Thomas and Znaniecki had found successors in a strict sense almost exclusively in Poland, where for the analysis of mostly written biographies special methodological tools (the biographical method) were developed<sup>1</sup>. Beside this, sociology mainly relied on anthropological traditions in dealing with the social construction and definition of different phases of the life cycle and the problems arising in the transition from one phase to the next<sup>2</sup>. In this tradition particular attention was given to childhood, youth and old age on the one hand, and to the family life cycle on the other hand<sup>3</sup>.

\* The plot routines, which are used in this paper, were first developed in the Nuffield Mobility Project of Oxford University in cooperation with B. H. S. Liénard, C. M. Llewellyn and John H. Goldthorpe. B. H. S. Liénard kindly made available to the author a version of the program implemented at Oxford and explained the details of the program structure. This program served as a model for the development of our own program which fits the particular structure of our life history data and the requirements of the CALCOMP plotter of the University of Mannheim. The Mannheim version corresponds in all essential characteristics to the Oxford version and makes use of some additional information. In cooperation with the author Gerhard Kolb of the Mannheim University computer center wrote the program. I wish to express my heartfelt thanks in particular, to B. H. S. Liénard and Gerhard Kolb, but also to Joan Kirchmeier who translated the text into English.

The paper forms a part of work being done in the SPES-Project of the Socio-Political Research Group at the universities of Frankfurt/Mannheim which is financed by the German Science Foundation.

<sup>1</sup> Thomas, William J., and Znaniecki, Florian, *The Polish Peasant in Europe and America*, New York 1919–1921.

<sup>2</sup> Eisenstadt, Samuel N., *From Generation to Generation: Age Groups and Social Structure*, Glencoe/Ill. 1956.

<sup>3</sup> Haller, Max, *Lebenszyklus und Familientheorie*. Bericht über das 8. Internationale Family Research Seminar in Paris 1973, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 26 (1974), pp. 148–166; König, René, *Soziologie der Familie*, in: König, René (ed.), *Handbuch*

It is hardly possible to bring the newly awakened interest in *life histories* onto a common denominator. However, common to most approaches is the observation that the standard methodology of empirical sociology in examining human existence in society on the basis of synchronic designs and cross-sectional data is unsatisfactory in many regards. Behavior and consciousness had to be regarded as the result of past life experience, and could therefore be explained only in the light of this process. A criticism of this kind particularly underlies symbolic interactionism and phenomenological analyses of the ‚Lebenswelt‘. With reference to this, for instance, Kohli explains crisis phenomena in mid-life as the result of anomic tensions between aspirations developed in the course of life and increasingly felt structural constraints which prevent these aspirations from being realized<sup>4</sup>.

But even authors who are not committed to this tradition have lately been stressing much more the importance of life-long experience and the need to include it in sociological research. Thus Bahrdt<sup>5</sup> gives special precedence to the study of life histories mainly for the analysis of consciousness structures among members of the lower classes, arguing that in these classes the relating of events constitutes the original form of reflection. „The articulation form used by members of lower-classes, including workers, when they reflect on their own situation . . . is the related story, the anecdotic account of their personal life history, especially of their childhood and youth.“<sup>6</sup> In accordance with this, Osterland<sup>7</sup> regards the investigation of sociobiographies as a decisive instrument with which to reconstruct the (political) socialization process and to analyze consciousness structures. That such analyses are bound up with a practically unsolvable problem is pointed out by Osterland himself. An unbiased picture cannot be portrayed, for „the individual perspective is always bound up with subsequent harmonization of conflicts, legitimation of decisions, repression and rationalization processes“<sup>8</sup>.

Leaving aside such difficult analyses of the genesis of consciousness structures, chiefly English authors consider the structure of life experiences to be a fundamen-

der empirischen Sozialforschung, Vol. 7, 2nd. ed., Stuttgart 1977.

Life cycle perspectives can clearly also be found in different traditions of psychology, be it in the psycho-analytical school, in development psychology or the newly expanding field of ‚life-span developmental psychology‘, where the questions of life-long personality development and the conceptualization of developmental phases and developmental crises prevail. Cf., e. g., Bühler, Charlotte, *Der menschliche Lebenslauf als psychologisches Problem*, Göttingen 1933; Erikson, E., *Identity, Youth and Crisis*, New York 1968; Nesselroade, John R., and Reese, Hayne W. (eds.), *Life-Span Development Psychology. Methodological Issues*, New York and London 1973.

<sup>4</sup> Kohli, Martin, *Lebenslauf und Lebensmitte*, in: *Kölner Zeitschrift für Soziologie und Sozialpsychologie*, 29 (1977), pp. 625–656.

<sup>5</sup> Bahrdt, Hans Paul, *Erzählte Lebensgeschichte von Arbeitern*, in: Osterland, Martin (ed.), *Arbeitssituation, Lebenslage und Konfliktpotential*, Frankfurt 1975.

<sup>6</sup> Op. cit., p. 14.

<sup>7</sup> Osterland, Martin, *Lebensgeschichtliche Erfahrung und gesellschaftliches Bewußtsein*, in: *Soziale Welt*, 24 (1973), pp. 409–417.

<sup>8</sup> Op. cit., p. 413.

tal factor in class-forming processes. Basically, the discussion here centers around the questions, in what way, to what extent and which types of intergenerational and intragenerational mobility processes can explain the evolution of class orientation or the absence of such orientation<sup>9</sup>. Particularly Goldthorpe/Llewellyn have made clear that only a differentiated investigation of life *histories* can make a profitable contribution to the study of class-forming processes<sup>10</sup>.

The present paper, however, is not concerned primarily with theoretical problems. The intention here is merely to present a methodical instrument which appears especially suitable for the description and analysis of life histories.

Life courses are processes which progress according to defined structures. In examining them, we aim to find out both the mode and extent to which they are structured, and the influencing factors more adequately and differentiatedly than everyday observations and common sense statements allow. Sociology's fundamental interest in life histories is rooted in the question of how the biological processes of growing up, maturing, aging and dying are linked with the „division du travail social“ and role allocations. An abundance of different solutions are conceivable and do in fact exist. The extent to which societies are stratified according to age was, it is true, regarded by e. g. Smith as an important criterion in distinguishing traditional and modern societies<sup>11</sup>. But even in modern societies not only familiar delimitations of specific groups of the population according to age and connected delimitations of life phases prove the relevance of age structures. The institutional affiliations of individuals and the transition processes between central institutions are clearly structured according to age, thus for example, the more or less total rooting in and subsequent withdrawal from the family of origin, the transition into educational institutions and from there into the occupational system, etc. In the occupational system biographical sequences are institutionalized in the form of professional careers which at least correlate with age.

There are time-tables which reflect social norms as to when certain events or role change should take place. But these time-tables do not as a rule display sharp dividing lines with general applicability. This is true at most for a few legal definitions such as the age of majority which, however, interestingly enough is of no particular significance for the subjective experience of life. In fact, it is merely an empirical question to determine how life courses are actually structured in relation to time;

<sup>9</sup> See Giddens, Anthony, *The Class Structure of the Advanced Societies*, London 1973; Goldthorpe, John H., *Mobilité sociale et intérêts sociaux*, in: *Sociologie et Sociétés*, 8 (1976), pp. 7–36; Goldthorpe, John H., and Llewellyn, Catriona, *Class Mobility in Modern Britain: Three Theses Examined*, in: *Sociology*, 11 (1977), pp. 257–287.

<sup>10</sup> Goldthorpe, John H., and Llewellyn, Catriona, *Class Mobility: Intergenerational and Work-life Patterns*, in: *The British Journal of Sociology*, 28 (1977), pp. 270–302. See also Bahrdt, *Lebensgeschichte*, and Bertaux, Daniel, *Mobilité sociale biographique: Une critique de l'approche transversale*, in: *Revue Française de Sociologie*, 15 (1974), pp. 329–362.

<sup>11</sup> Smith, Michael G., *Pre-Industrial Stratification Systems*, in: Smelser, Neil J., and Lipset, Seymour M. (eds.), *Social Structure and Mobility in Economic Development*, Chicago 1966.

what are typical sequences and how strongly they are adhered to; what is the duration of single phases of sequences; in what way different sequences are co-ordinated; how the fact of being too early or late in one sequence influences the course of another sequence; whether delays are made up for or a premature advantage disappears with time. In addition, a wide field of questions concerns itself with the differences in life patterns in various population segments of the social structure<sup>12</sup>.

It must be emphasized that age as such is not regarded here in any way as a causal variable. It is social norms which make age in some ways appear to be a good indicator variable. These social norms are linked partly explicitly with age, but partly only implicitly on account of the fact that certain sequences are normed. Both explicit norms and regularities which result from particular sequences (e. g. from replacement processes) can change with time. In analytical models it should therefore be ensured that this variation in time can be accounted for.

From these reflections arise a number of requirements which must be borne in mind when analyzing life histories. The first problem lies in the fact that we are dealing with longitudinal data. These cross-time data, which can cover whole or partial life histories, are of interest not just from the point of view of how particular states (e. g. occupation, family situation or attitudes) alter between two or more observation points. In addition important questions relate to the duration of single states and their consecutive sequence.

The second problem results from the fact that states, occurrences and developments in different spheres of life must be inter-relatable, as far as possible, both in their duration and their sequence. Furthermore, these inter-relations cannot be regarded in isolation for particular individuals only, but, at least for certain phases of life, groups of individuals (married couples, families) form the units of analysis, and the model should ensure that the experiences of the individual members of these groups can be related one to another.

If we disregard the primarily Polish investigations based on the biographical method, and the mainly qualitative ethnomethodological studies on status passage, in mobility research we find at most approaches which are connected with the analysis of life histories<sup>13</sup>. However, by comparison to the program just outlined, these approaches are extremely modest. Basically, two research paradigms are present in mobility research<sup>14</sup>.

The first of these relates to the investigation of transitions between a set of positions within certain time intervals (e. g. transitions from educational levels to occupational positions or between occupational positions from one point of obser-

<sup>12</sup> Compare here: Clausen, John A., *The Life Course of Individuals*, in: Riley, Mathilda W., et al. (eds.), *Aging and Society*, New York 1972.

<sup>13</sup> E.g. Strauss, Anselm L., *The Context of Social Mobility. Ideology and Theory*, Chicago 1971.

<sup>14</sup> See Sørensen, Aage B., *Models of Social Mobility. Report No. 98*. The John Hopkins University, Baltimore 1971.

vation to another). In this analysis of transition matrices the simultaneous inclusion of several points of observation is doomed to rapid failure due to problems of sample size. If a somewhat differentiated classification of positions is used, this limit will quickly be reached when, for example, background, education and occupational position are examined in a transition matrix model<sup>15</sup>. If, moreover, a sufficiently differentiated description just of occupational careers is to be made, such a large number of transition matrices will be needed that clarity will inevitably be lost. The use of complex Markov chain models generally fails, due to the unrealistic assumption of constant transition probabilities irrespective of earlier moves<sup>16</sup>.

The second paradigm comprises regression analyses of the status attainment process. Here the number of considered variables can easily be increased. However, in the simple form of path analyses, as developed by Blau and Duncan<sup>17</sup>, this model has, along with statistical problems considerable theoretical limitations as pointed out primarily by Crowder<sup>18</sup>. In respect to both paradigms, Sørensen's contributions represent important steps forward<sup>19</sup>. Since, however, the approaches developed in mobility research are basically limited to the analysis of occupational careers, the problem of simultaneously taking into account events and developments in various other spheres of life remains unsolved.

A wholly different approach to the analysis of life courses is contained in the graphs first developed by Form and Miller also for the investigation of occupational careers<sup>20</sup>. Miller and Form drew graphs of the occupational careers of all respondents in a small sample within fairly rough occupational classifications, and illustrated from them the typology of various phases of working life (initial, trial, stable work period). In the last few years Carr-Hill and MacDonald and Lienard and Llewellyn have adopted this approach and, by developing EDP plot routines, have

<sup>15</sup> Müller, Walter, and Mayer, Karl Ulrich, *Chancengleichheit durch Bildung? Untersuchungen über den Zusammenhang von Ausbildungsabschlüssen und Berufsstatus* (Gutachten und Studien der Bildungskommission des Deutschen Bildungsrates, vol. 42), Stuttgart 1976.

<sup>16</sup> The use of Markov chain models for the study of mobility processes is discussed by McFarland, David D., *Substantive Contributions of Markov Models to the Study of Social Mobility*, University of Chicago 1974, mimeo.

<sup>17</sup> Blau, Peter M., and Duncan, Otis D., *The American Occupational Structure*, New York 1967.

<sup>18</sup> See Müller, Walter, *Familie, Schule, Beruf: Analysen zur sozialen Mobilität und Statuszuweisung in der Bundesrepublik*, Opladen 1975; and Müller, Walter, *Education and Class Structure in West-Germany* (SPES-Arbeitspapier Nr. 32), Frankfurt/Mannheim 1975; Crowder, David N., *A Critique of Duncan's Stratification Research*, in: *Sociology*, 8 (1974), pp. 19–45.

<sup>19</sup> Sørensen, Aage B., *A Model for Occupational Careers*, in: *American Journal of Sociology*, 80, 1 (1974), pp. 44–57; *Growth in Occupational Achievement. Social Mobility or Investments in Human Capital*, in: Land, Kenneth C., and Spilerman, Seymour (eds.), *Social Indicator Models*, New York 1975; *The Organization of Activities in Time*, in: *Demography*, 19 (Aug. 1975); *The Structure of Intragenerational Mobility*, in: Ziegler, Rolf (ed.), *Anwendung mathematischer Verfahren zur Analyse des Statuszuweisungsprozesses*, Kiel 1975.

<sup>20</sup> Form, William H., and Miller, Delbert C., *Occupational Career Pattern as a Sociological Instrument*, in: *American Journal of Sociology*, 54 (1948), pp. 317–329.

created the possibility of exploiting this procedure with much greater ease and variety<sup>21</sup>.

As against Form and Miller's hand-drawn graphs, sub-groups for single diagrams can not only be picked out at will and then compared very easily, but a great step forward is also represented by the fact that, apart from the occupational patterns which Miller and Form restricted themselves to, a large number of other relevant events in life histories can be considered. Lastly, linking with a data retrieval system makes it possible to standardize the life history of different individuals in relation to various events or temporal orientation points. Thus life histories can readily be plotted in historical time, i. e. according to the calendar years in which they occurred, on the one hand, but also according to the age of the individuals concerned, on the other. The standardization can also be carried out according to any other event considered relevant, for example, the year of entry into working life, or for women the year of marriage or birth of the first or last child. Depending on the underlying question posed, various aspects of life courses can thus be easily emphasized on the graphs.

In the following I shall present results of first attempts to plot life histories by means of such EDP plot programs and to utilize such graphs for analytical purposes. First of all, though, I shall describe briefly the data I will be using.

The data are taken from the microcensus supplement of the Federal Republic of Germany of April 1971 („Social and Occupational Mobility of the Population“). Going beyond the basic program of the annual microcensus of the Federal Statistical Office, various information on the social background of the respondents, their education, and particularly the course of their occupational career, was gathered for a selection of 1 % of the population. Furthermore, data are available on the year of marriage, the number of children according to age groups, and change of address between 1960 and 1971. The data on occupational careers relate to the occupational position and branch of the economy in the years 1939, 1950, 1960, and 1971. In addition, all changes in occupational position and branch of the economy are recorded for the time period from 1960 to 1971, so that — at least in so far as the data are correct — a year by year analysis of occupational careers can be undertaken for this period. Since all members of a selected household were questioned in the microcensus, information on all household members is in principle available and can be combined. In the following illustrations the starting population consists of men. If they were married and living together with their wives, their data was linked with that of their wives, so that the life courses of married couples could be plotted jointly.

The present data were, of course, not collected from the standpoint from which they are now being analyzed. Therefore, substantial information which would have been useful for our analysis is missing. Thus we would, for example, have wished

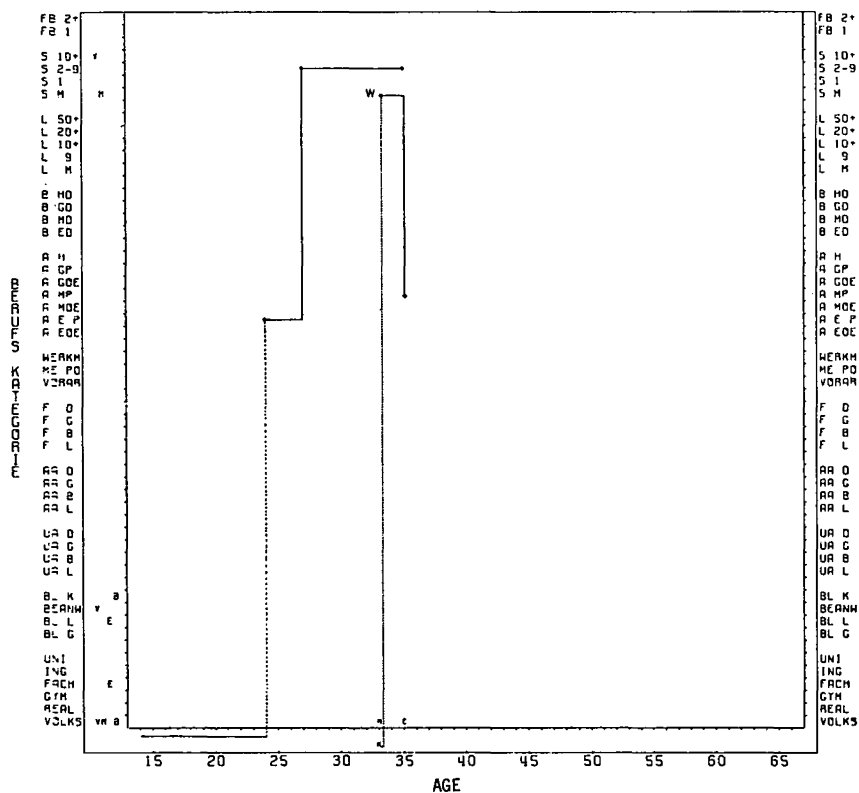
<sup>21</sup> Carr-Hill, R. A., and MacDonald, K. I., Problems in the Analysis of Life Histories, in: Sociological Review Monograph, 19 (1973), pp. 57–95; Liénard, B. H. S., and Llewellyn, L. M., The Analysis of Life History Data Using Graphical Devices (no year), mimeo.



that the entire educational and occupational career were recorded in detail, or that information on the moment of exit from the parental household, on change of address, exact data on the time of birth of children, on phases of further education, illness, military service, imprisonment, etc., were available.

In the following presentation, we must take into consideration the limitations arising from the data material available to us. In particular, the illustrations can essentially refer only to occupational careers, since information on other spheres of life is insufficient or non-existent. However, it will become evident that we could easily make use of the procedure also for non-occupational information. Let us take an example to illustrate the nature of the graph (see Figure 1):

Figure 1: Example of a Single Life History Plot



The continuous horizontal lines within the framework of Figure 1 symbolize periods spent in different branches of the educational system or in occupational positions indicated by the labels outside the framework of the Figure (the meaning of the different labels is explained in Appendix 1). The vertical lines portray changes in these positions. Changes in occupational position (e. g. change of firm, so far as they are included in the data), which do not signify a change in the categories under consideration, are indicated by a rhomboid identification symbol on a horizontal line. Changes resulting in a state not included in the defined educational and occupational categories (e. g. no response, military service, illness, unemployment) are symbolized by dotted vertical lines and lead out of the inner framework. In the space between the two lower limits of the framework a symbol is inserted in these cases to signal the state in question. For the time being, the following symbols have been allotted:

- A = unemployed
- B = military service
- H = housewife
- K = no response

Keeping within the framework, further symbols can be situated on the lower horizontal axis. They represent the following events for the respective year:

- C = birth of a child
- D = divorce
- L = change of address
- M = marriage
- Q = exit from parental household

The first column of the graph can also contain symbols. They indicate what school education and occupational training the respondent (B), his wife (E), his father (V) and his mother (M) have. For the father and mother, their occupational position is also indicated.

Within the framework itself not only the career-line of the respondent is drawn, but from the moment of his marriage also that of his wife. For purpose of distinction, the wife's line is marked with a „W“.

Hence, Figure 1 relates the following life history of a respondent who was 35 years of age in 1971: The respondent comes from a family in which both parents have only primary school education. In addition, the father completed a civil service preparatory course, but then became self-employed and owned a business with more than 10 employees. The mother was occupied in the business as a co-working family member. The respondent similarly had only primary school education, but did a commercial apprenticeship. Up to the age of 23 there is no information on his occupational career. From the age of 24 to 27 he is a lower white-collar worker in the private sector. Then he sets up on his own and has a smaller firm with 2—9 employees. At 33 he marries a woman who has completed technical college and a household apprenticeship. At first his wife works in her husband's business as a co-

working family member and thereafter becomes a middle white collar worker in the private sector. She has a child in the same year. The respondent has lived in the same place since the age of 15.

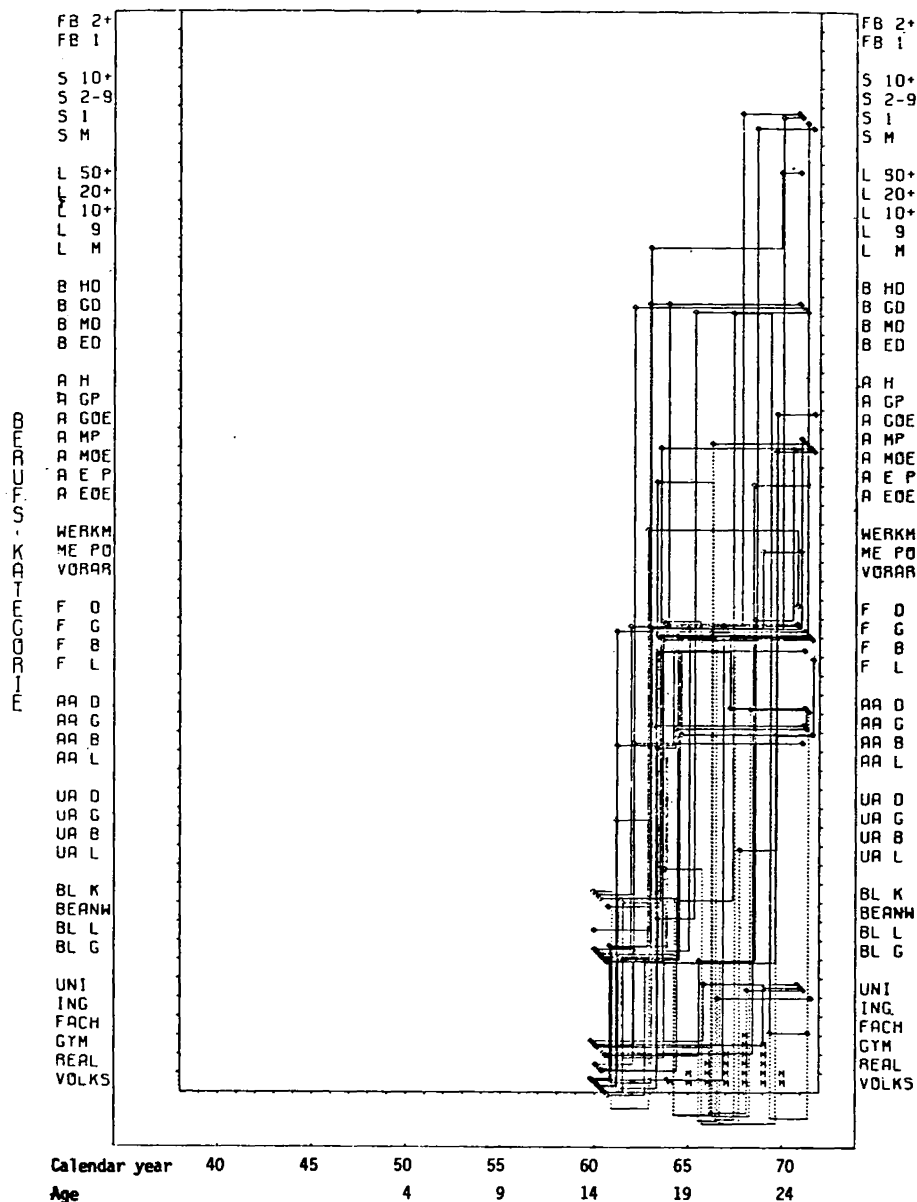
It would easily be possible to include other data in this kind of graph by means of symbols or another form of path curve, so that eventually numerous events and developments in a life history can be portrayed in time with relative economy of space and at a glance. In principle, there is no reason why subjective experiences and judgements, as long as they can be pinpointed in time, should not be included in this type of graph.

It would be difficult to compare large numbers of single life plots in order to illustrate and recognize certain structures of life courses. Therefore, the program provides the possibility of representing several life courses at the same time in one graph. The line for each successive person is set laterally a short distance away from the previous one, so different life course patterns can be distinguished. On the other hand the number of lines running parallel to each other express the quantitative importance of a given pattern and certain structural elements are emphasized on the graph in the form of thick bars.

Figures 2–5 each contain such multiple plots for birth cohorts of 10-year intervals. Each diagram contains the occupational careers of 40 men selected at random. For clarity the wife's career, data on the birth of children, change of address and background have been excluded. When looking through these diagrams, the structure of the available data must once again be recalled. Only for the period 1960 to 1971 are yearly data available. The previous years are represented by observations for the sample years 1939 and 1950 only. Therefore, mobility between the various occupational categories is also registered only for these points in time. Thus, for instance, the entire period 1950–1960 is represented by the situation in 1950 and in 1960 and only the most recent change is accounted for. Supposing, for example, a change took place in 1954, but in 1959 the person returned to the starting category of 1950, these two changes will not be registered here at all. This is, of course, extremely unsatisfactory. However, if this is a cause for objection it should be borne in mind that in traditional analyses of life course data (by means of transition matrices or regression analysis) as a general rule only information for particular sample years of fixed points of observation are taken into account. The problematic nature of such selective observations appears rather drastic here because they contrast with the yearly observations and portrayals during the period 1960–1971.

First of all let us consider the youngest cohort, members of which are 14–25 years of age in the period 1960–1971. This phase of life is marked by the transition from educational institutions to the occupational system. In the year 1960 a fair majority of the 14-year-olds is doing an industrial, trade or commercial apprenticeship. In the following year they are joined by a few more who in 1960 were still in primary school, and even later by a few secondary and upper school-leavers. The number of those embarking on a commercial apprenticeship is much lower than that for an industrial apprenticeship. Most gymnasium students stay at school until approximately the age of 20 and then go to university, some of them not until they

Figure 2: Occupational Life Histories of 40 Randomly Selected Men of Birth Cohort 1946



have done a few years military service. (The period of military service is portrayed by the broken lines running in the direction of the lowest part of the diagram.)

The first to take up an occupation are employed as unskilled or semi-skilled workers. Their number is at first fairly small, but later on they are joined by some who have completed an apprenticeship or worked previously as skilled workers. Almost all of those with a trade or an industrial apprenticeship — some after a break for military service — become skilled workers. Most of them are still skilled workers at 25. A few become white-collar workers in middle positions towards the end of the period of observation. But at least as many then carry out semi-skilled work. Only in odd cases do skilled workers rise to supervising positions or become self-employed before the age of 25. Little can be said here about the occupational future of commercial apprentices because the lines in this diagram cannot be clearly followed. At a later point I shall talk about a possibility of solving this descriptive problem.

Altogether — with the exception of the skilled worker area — the structure of Figure 2 is primarily vertical in character. This means that the respondents undergo many changes during the period of investigation. If we now compare the graph for the birth cohort 1946 with the diagrams of the older birth cohorts (1936, 1926, 1916), we notice that for the period 1960–1971 the structure of the graphs shifts more from the vertical to the horizontal. For the periods 1939–1950 and 1950–1960 the horizontal lines are in any case affected by the structure of the data, which, of course, register changes only for the sample years. With increasing age of respondents, however, vertical movements also decrease in the period 1960–1971, for which yearly observations are available. Hence the data clearly confirm the postulate of decreasing mobility rates with increasing age. Between the ages of 25 and 35 (see the last period in Figure 3) more cases of change take place than between the ages of 35 and 45 (see the corresponding period in Figure 4). Mobility decreases further between the ages of 45 and 55. However, the data also show that even after the age of 40 mobility processes do still occur. This is also true for workers, for whom — in all skepticism towards the classification employed — we observe both upward and downward mobility.

Finally a further interesting observation arises in the temporal pinpointing of the mobility processes observed between 1960 and 1971. If mobility processes in general were to decrease with age, most movements would be expected to take place at the beginning of the respective period of observation. But according to Figures 4 and 5 for the two older cohorts, more movements occur in the second half of the period 1960–1971, so that for further investigations the question arises, to what extent are effects of age linked with effects of historical periods<sup>22</sup>.

One problem with multiple plots is that the graphs become too complicated with larger numbers of people. If the number is too large, it is no longer possible to identify clearly to which observation year and category a particular point belongs. Since each subsequent line is placed minimally downwards and to the right of the previous one, the area provided for individual categories and observation years can be

<sup>22</sup> In more detail this problem is discussed in Müller, *Class Structure*.

**Figure 3: Occupational Life Histories of 40 Randomly Selected Men of Birth Cohort 1936**

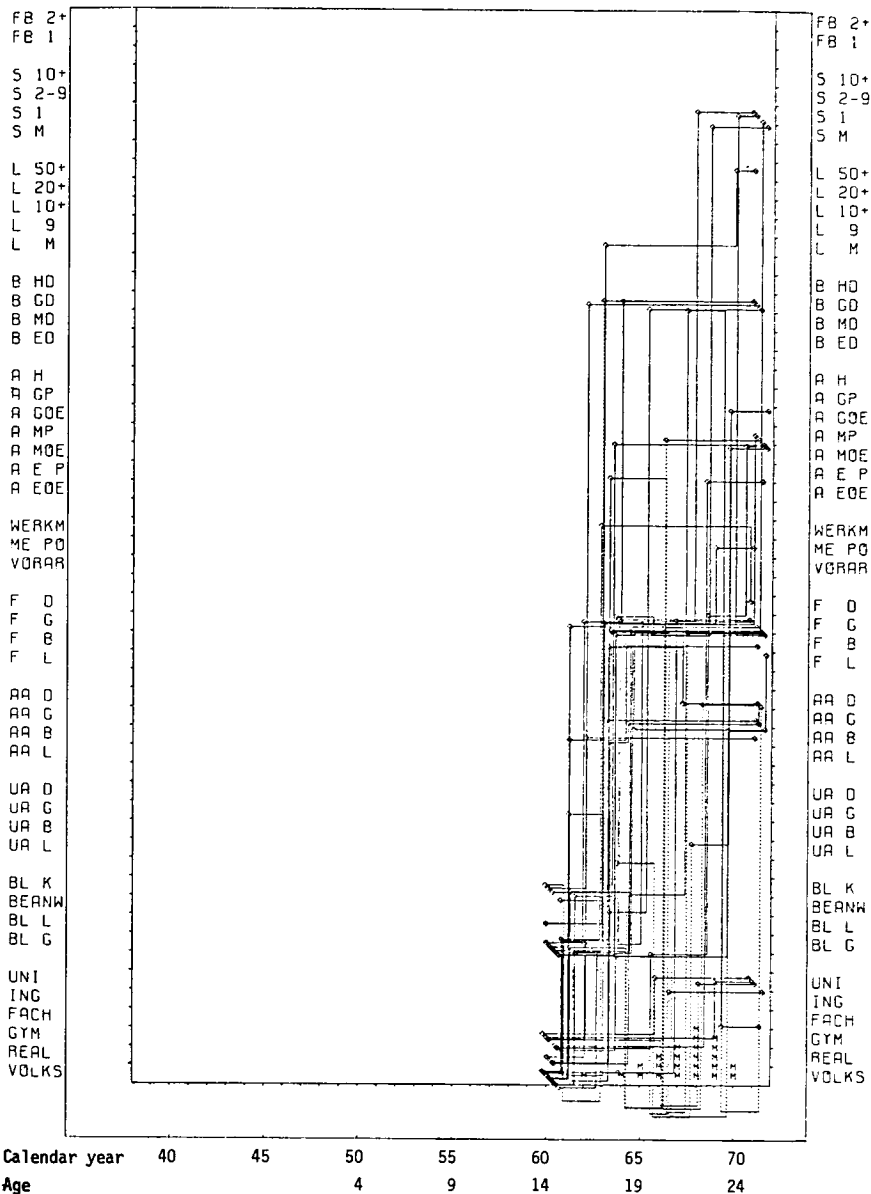


Figure 4: Occupational Life Histories of 40 Randomly Selected Men of Birth Cohort 1926

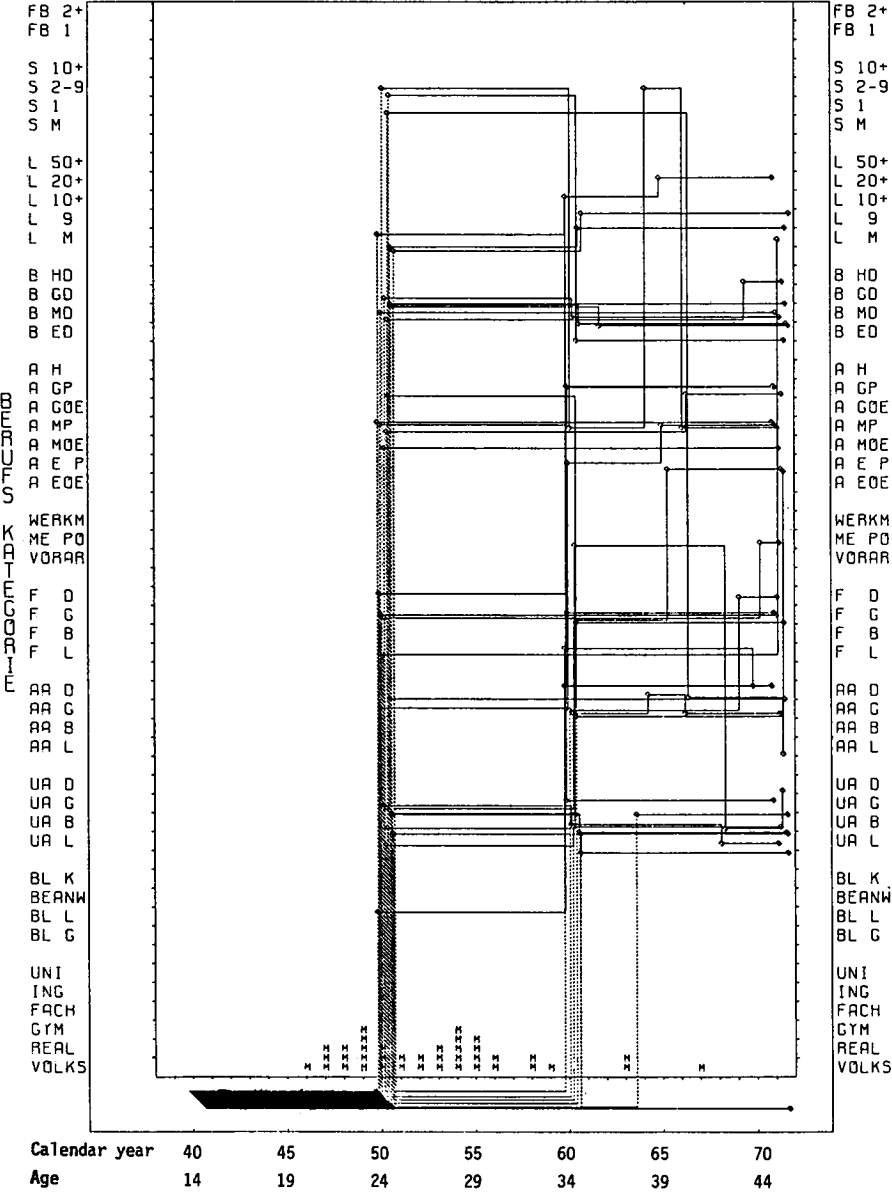
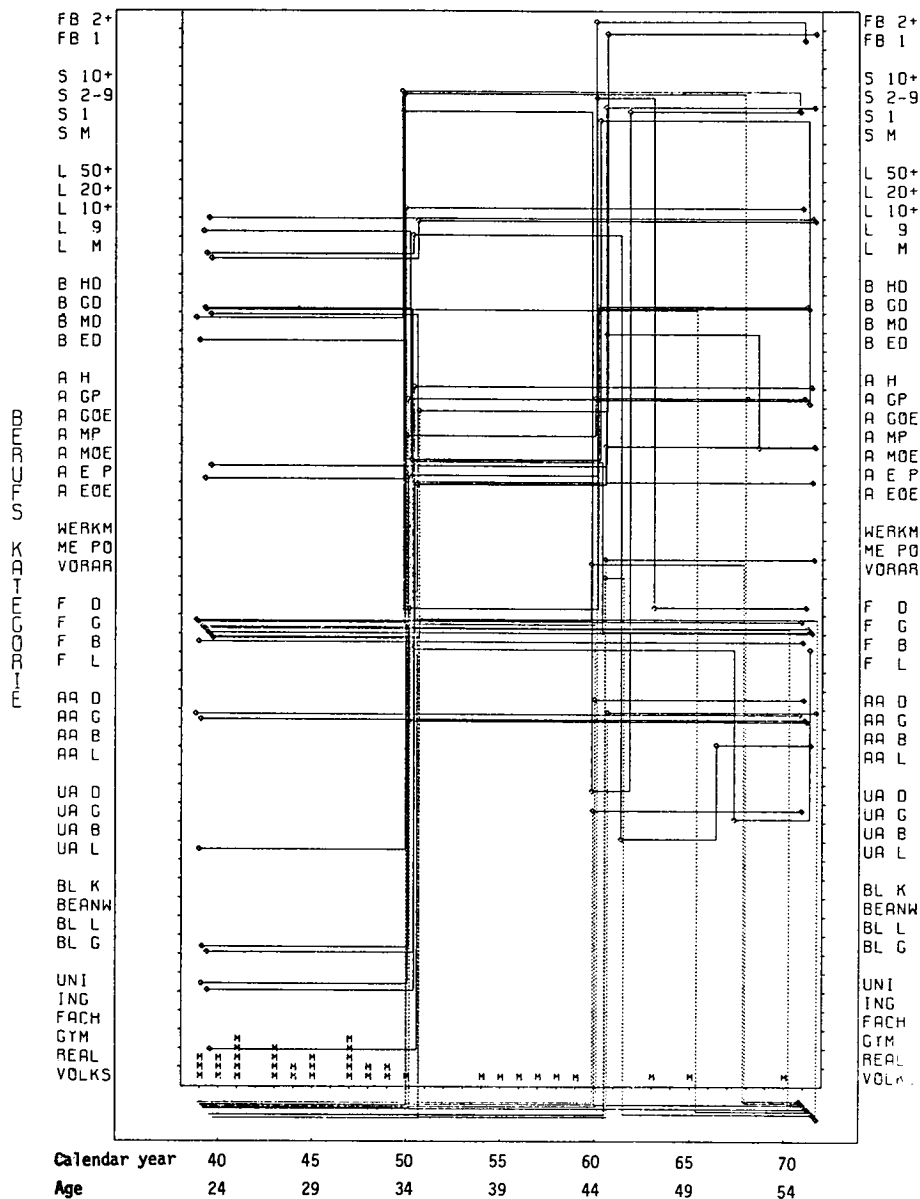


Figure 5: Occupational Life Histories of 40 Randomly Selected Men of Birth Cohort 1916





exceeded. Choosing a larger format is only a limited help. This is certainly a possible way out for the researcher during the course of his work, but it is hardly realistic from the point of view of publishing results. There remain, however, two other possibilities. First, the data can be divided up at random onto different plots by means of the split half procedure or even more subdivisions. This method is based on the principle of replication of results through mutually independent observations. The second method involves systematically selected observed cases according to theoretical criteria and portraying these various subgroups separately. By this means, the maximum possible number of persons can be portrayed for each sub-group and in this way the number of cases investigated increased. In combination with tabular classifications, the relative importance of sub-groups can be determined. The subdivision into birth cohorts employed previously was basically such a sub-group formation. But the sub-group formation model lends itself especially either to isolating out quantitatively important groups (typical cases) and examining them in more detail, or to forming sub-groups which, for theoretical reasons, one can expect to contrast. In the following I should like to give a few illustrations for both cases.

Quantitatively, the most significant groups in the transition from the educational to the occupational systems are those who have completed occupational apprenticeships. For the long-term occupational future, it makes a pronounced difference whether this was a trade-industrial or commercial apprenticeship. In Figures 6 and 7, 30 people with completed apprenticeships for each of the two types are compared. At the same time of observation all respondents were 35 years old. The phase of which yearly information is available relates to the respondent's age of 24–35. The first time of registration, the year 1950, happens to be their 14th year, at which point the majority of the future apprentices were still at school. Only one of the commercial apprentices but about one quarter of the industrial apprentices had already started their apprenticeship. The on average higher level of education among the commercial apprentices is also very apparent.

At 24 years of age a good fifth of the commercial apprentices were still white-collar workers in lower positions. But before their 30th year all except two had climbed to middle positions. From middle positions occupied by almost half of the respondents at 24 years of age, a few „descend“ to lower positions, a tendency which is countered later on. Otherwise, some climb to higher positions, so that at the end of the period of observation the proportions of middle and higher positions are balanced and amount to approximately one third. The career stability of the commercial apprentices who become civil servants is remarkable. Not only did no single person who became a civil servant between the ages of 24 and 35 lose that status — which is in itself hardly surprising — but not one of them even changed his career group in this time<sup>23</sup>.

Compared with the commercial apprentices, the occupational career of the industrial apprentices appears partly much more stable and partly much more

<sup>23</sup> The basis of observation is however very small. A special analysis with a sample of civil servants would be necessary here.

Figure 6: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Which Have a Commercial Apprenticeship

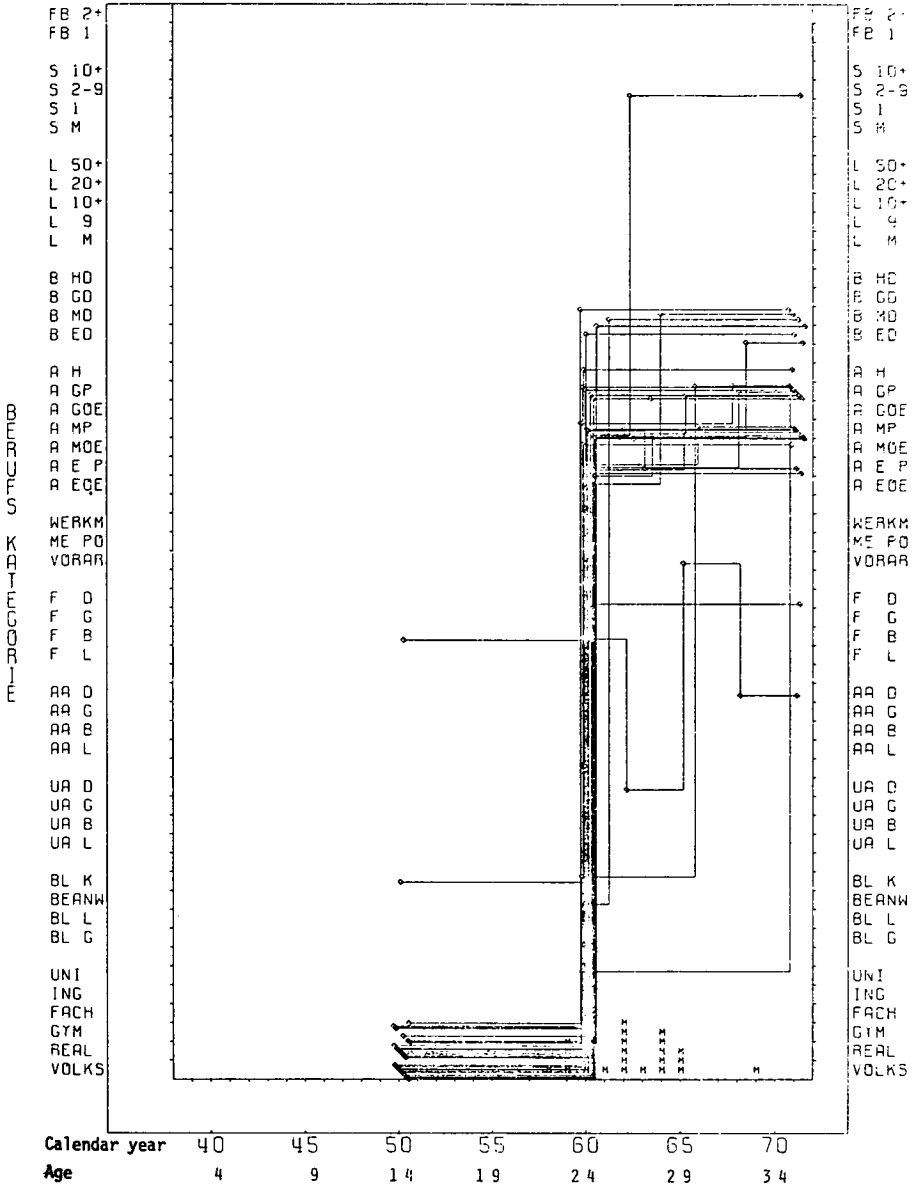
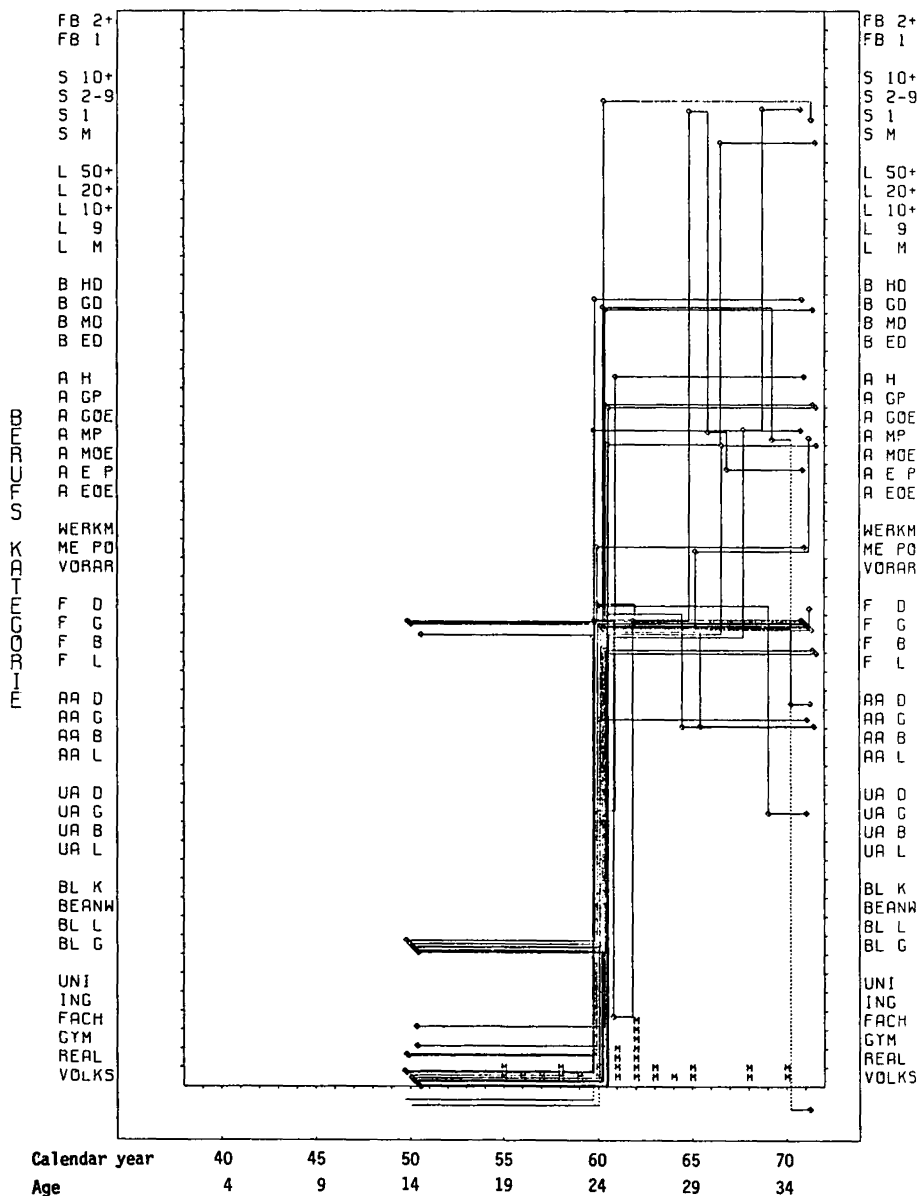


Figure 7: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Which Have a Trade or Industrial Apprenticeship



uncertain. About half of all those included do not change their occupational position at all between the ages of 24 and 35. They remain skilled workers the whole time. However, of those who once changed from the position of skilled workers, none reverted back. This holds true both for those who transfer to semi-skilled or unskilled work and for those who attain supervisory positions, become white-collar workers, civil servants or self-employed. If differentiated data for the period from 1950 to 1960 were available, it would be possible to clarify to what extent former industrial apprentices, who are higher white-collar workers or civil servants in 1971, attained this climb through successive occupational improvements or through interim education such as at an engineering school or technical college. Higher civil servants have in any case probably taken the latter course because of the existing formal regulations regarding entry.

The career pattern of the trade and industrial apprentices differs also from that of the commercial apprentices in that irregular careers are no exception<sup>24</sup>. This is true not only of skilled workers who carry out less qualified work later on. An example of an irregular career is, for instance, the case of a skilled worker who goes into self-employment, a year later takes on a middle white-collar position and a year after that a lower white-collar position in which he remains. A further example is that of a skilled worker who becomes an upper-grade civil servant, then a white-collar worker in a middle position and finally unemployed.

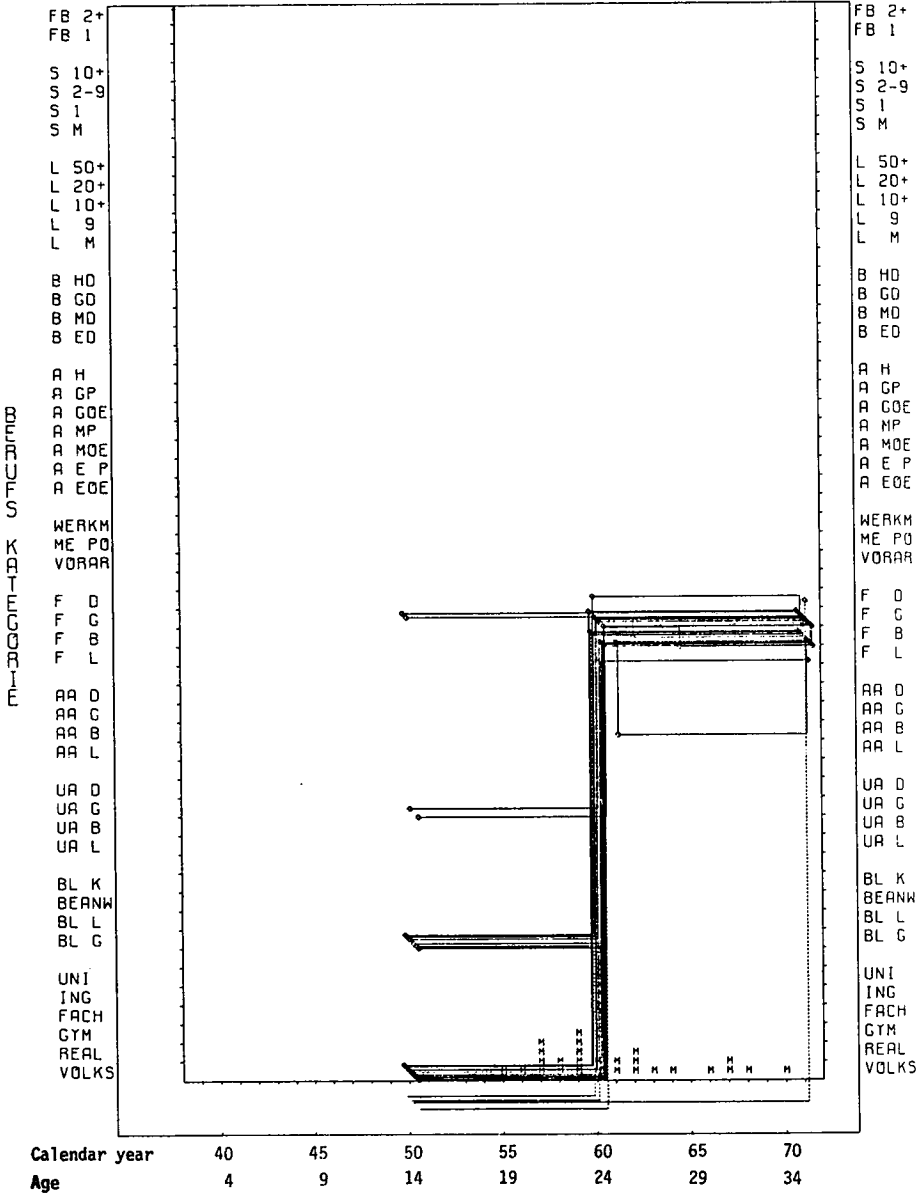
A further interesting difference between the groups relates to the extra-occupational sphere. The marriage age of the former industrial apprentices is much more scattered than that of the commercial apprentices. In principle it would be no problem to investigate the incidence of further extra-occupational events in the same way as for marriage, if data were available for them.

Sub-groups can thus be traced, as we have just seen, from a given starting point on into the future. But it is also justified to ask what kind of life path have those people had who have all attained a particular goal? In the following emphasis will be placed on this aspect of viewing life histories. Figure 8 shows the occupational past of a group of people picked at random, who at 35 years of age are all skilled workers. The result is very clear. Nearly all these skilled workers have worked in that capacity without interruption between the ages of 24 and 35. In all probability, only very few of them have had a break from skilled work before that — with the exception of time spent doing military service.

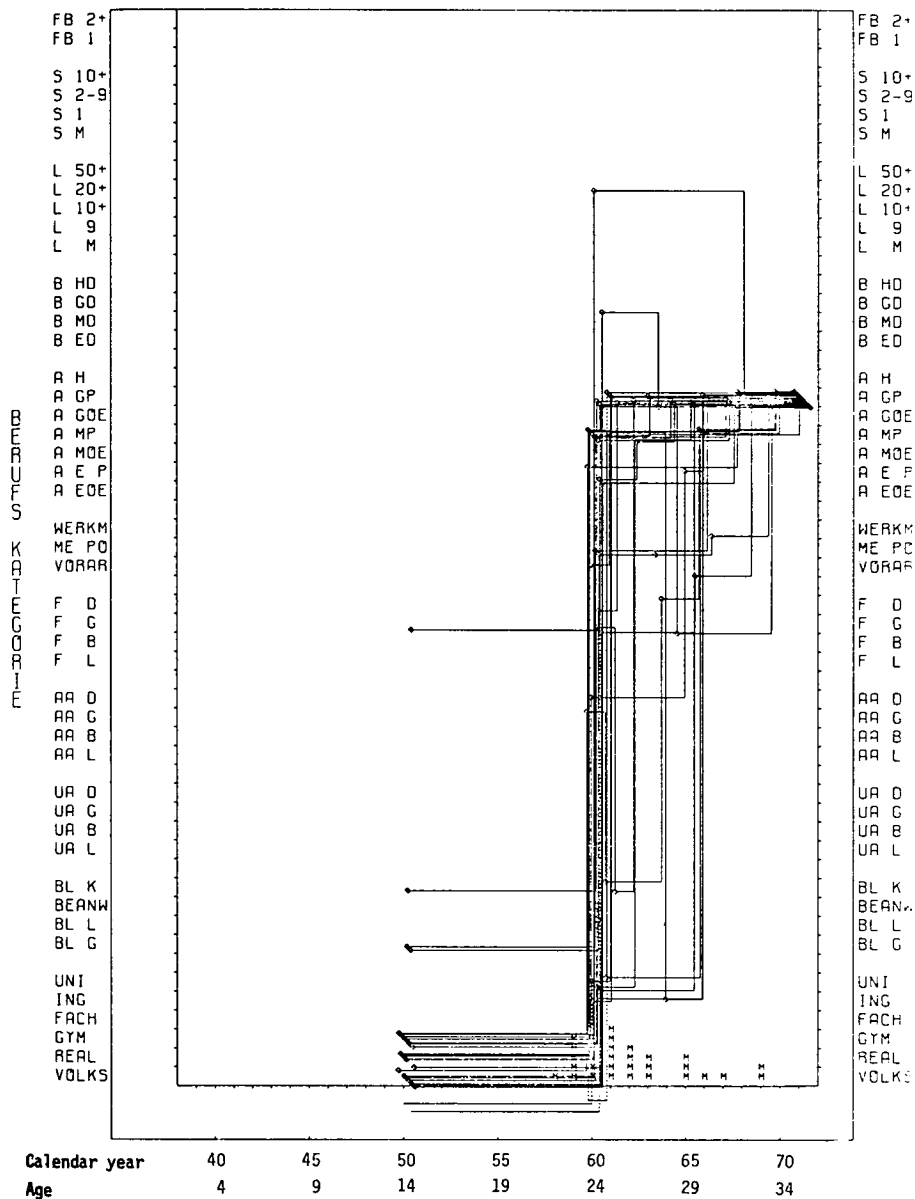
On the other hand — in order to pick out another group — completely different and much more heterogeneous is the occupational experience of those who at 35 years of age occupy higher white-collar positions (see Figure 9). Those who enter these positions directly from university and only have a few years working experience are to be found in this group. But the majority have attained these positions by means

<sup>24</sup> Wilensky, Harald L., *Orderly Careers and Social Participation: The Impact of Work History and Social Integration in the Middle Mass*, in: *American Sociological Review*, 26 (Aug. 1961), pp. 521–539;

**Figure 8: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Who Are Skilled Workers at Age 35**



**Figure 9: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Who Are Lower Managers at Age 35**



of successive occupational rises which resemble the rungs of a ladder. If lower and middle white-collar positions predominate among the previous positions held, occasionally even former unskilled workers make their way to higher white-collar positions via successive rises.

In a different context the hypothesis was suggested that these different access paths are closely linked with the social background of the future higher white-collar workers<sup>25</sup>. Middle class families make use of the educational system in order to guarantee a profitable occupational future for their children. When children from working class families manage to attain higher occupational positions, then this is more likely to occur via successive occupational rises. A test of this assumption is to be undertaken in Figures 10 and 11. In each Figure 30 men picked at random, all of whom occupied higher white-collar positions at 35, are contrasted. Those in Figure 10 come from working class families and those in Figure 11 from higher manager and civil servant families.

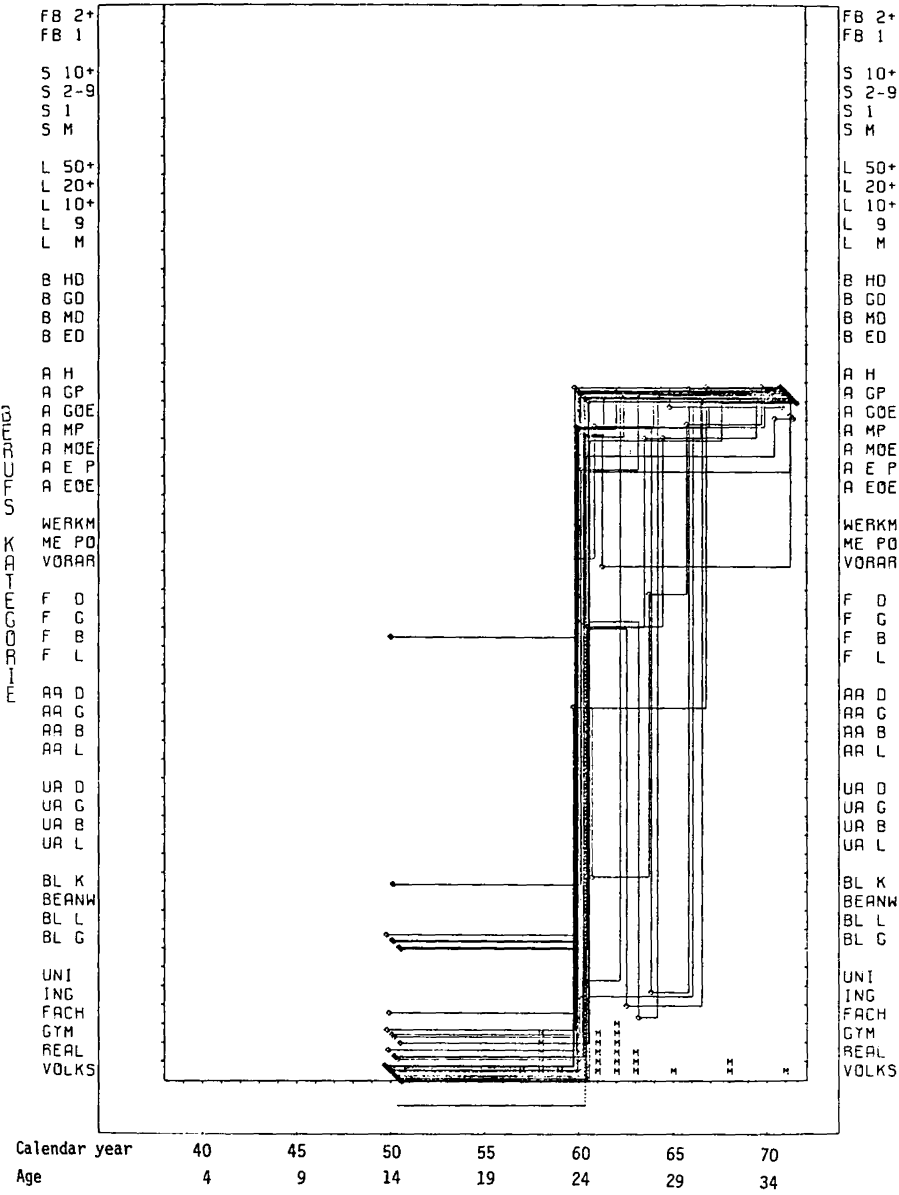
Even the distribution within the school system which these two groups of respondents display at the age of 14, shows marked differences. The children of workers are in the vast majority at primary school or doing an apprenticeship. Middle class children, on the other hand, are by a similar majority attending gymnasium, secondary school or technical college. Unfortunately, their further path of development up to the age of 24 cannot be traced in detail. But even the differences between the two groups which appear after that are clear enough. The majority of middle class sons already occupy higher positions or are still at university. From university most of them then transfer directly to higher occupations before the age of 30. Only about one quarter is working at 25 in lower or middle white-collar jobs and attains higher positions over the years. In contrast, only a small number of workers' sons of this age are already in higher positions or still at university. Most of them are in middle or lower jobs, are occupied both as white-collar and blue-collar workers and only move up into higher positions over the years. This is achieved partly through intermediate phases of training, particularly by attending technical and engineering colleges<sup>26</sup>. Altogether, the two diagrams document very clearly the influence of social background on the path of access to higher occupational positions.

As a final illustration of the possibilities of analysis of life patterns by means of graphs, we take reference to Figures 12 and 13. These diagrams relate to the wives of the men represented in Figures 10 and 11. Whereas previously the wives were excluded from the plot, now their husbands are excluded. Here, too, there are obvious differences between the two groups. First of all — and this is also visible in Figures 10 and 11 — the time of marriage in the group with working class husbands is earlier

<sup>25</sup> Müller, Walter, *Klassenlagen und soziale Lagen in der Bundesrepublik*, in: Handl, Johann, et al., *Klassenlagen und Sozialstruktur. Empirische Untersuchungen für die Bundesrepublik Deutschland*, Frankfurt 1977.

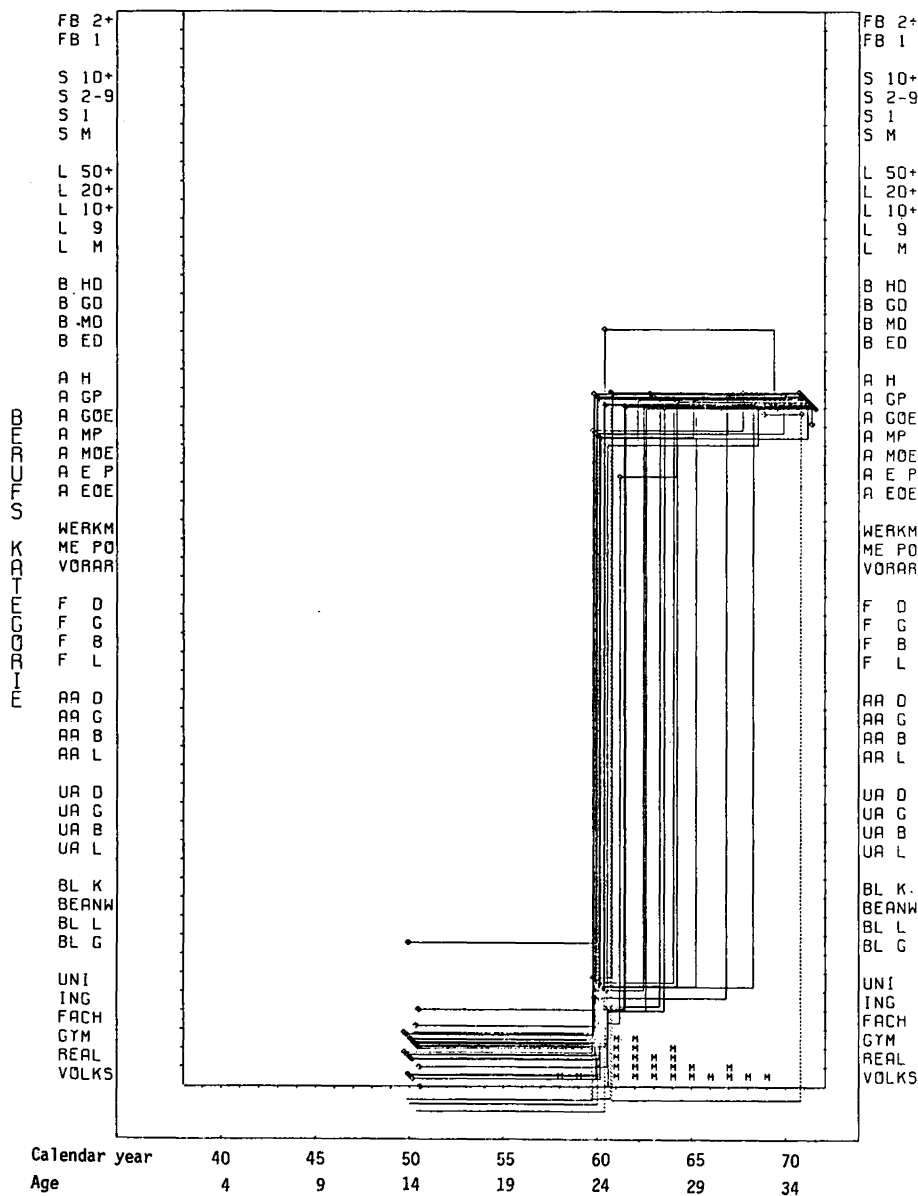
<sup>26</sup> For the significance of this type of educational institution as a means of attaining higher positions, see also Lutz, B., and Kammerer, G., *Das Ende des graduierten Ingenieurs? Eine empirische Analyse unerwarteter Nebenfolgen der Bildungsexpansion*, Frankfurt 1975.

Figure 10: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Who Are Lower Managers at Age 35 and Whose Father Was a Worker

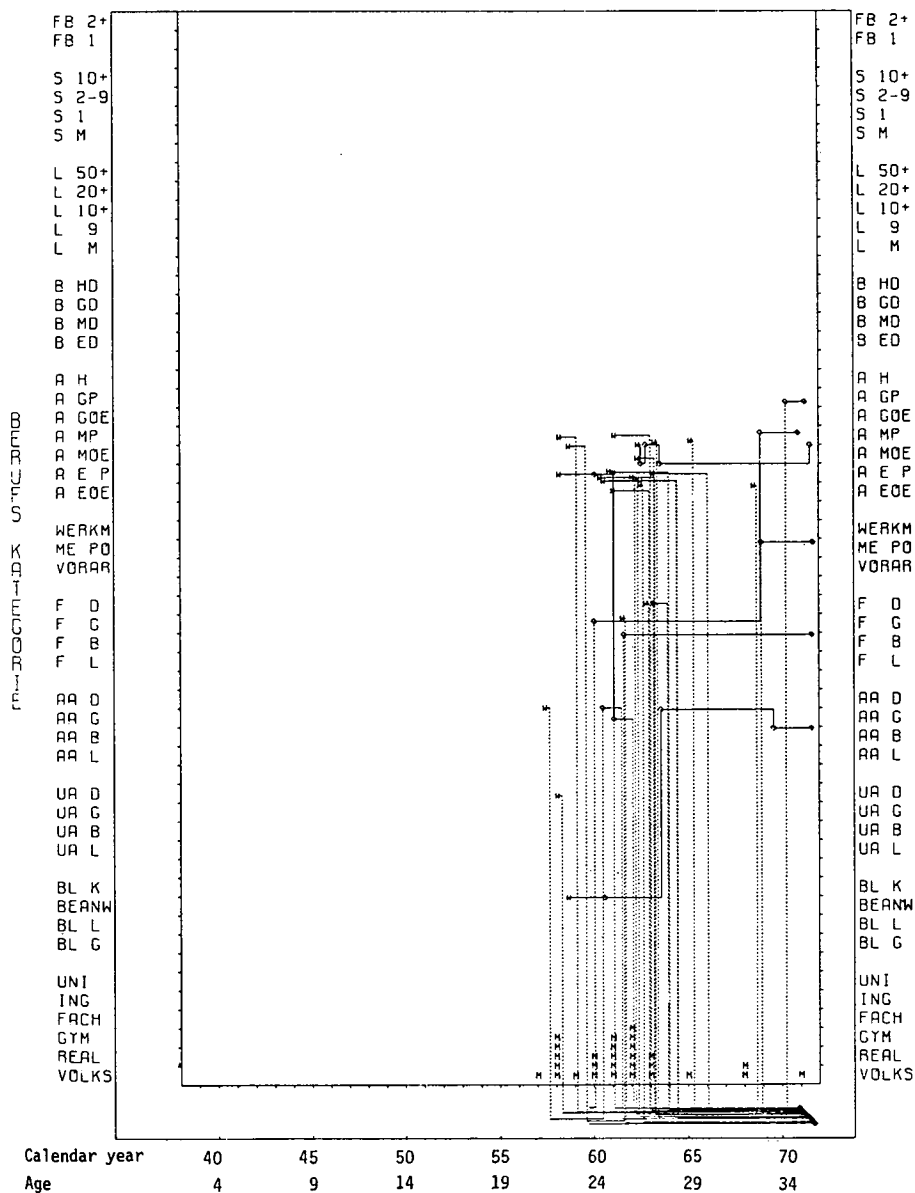




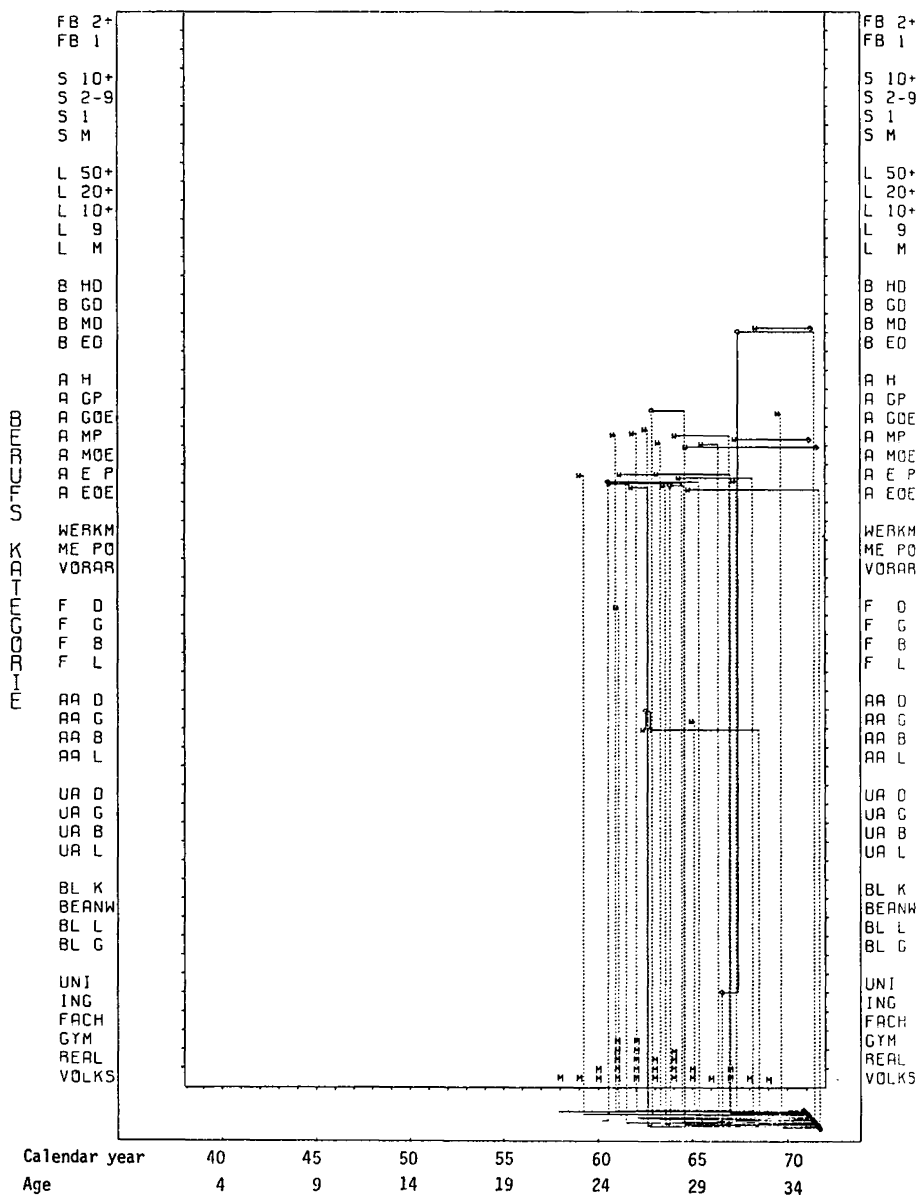
**Figure 11: Occupational Life Histories of 30 Randomly Selected Men of Birth Cohort 1936 Who Are Lower Managers at Age 35 and Whose Father Was a Lower or Top Manager or Upper High Trade Civil Servant**



**Figure 12: Occupational Life Histories of the Wives of 30 Selected Men of Birth Cohort 1936 Who Are Lower Managers at Age 35 and Whose Father Was a Worker**



**Figure 13: Occupational Life Histories of the Wives of 30 Selected Men of Birth Cohort 1936 Who Are Lower Managers at Age 35 and Whose Father Was a Lower or Top Manager or Upper or High Trade Civil Servant**



than in the middle class group. This is certainly connected with the later entry into working life in the middle class group. Secondly, the occupation carried out by women of the middle class group before their marriage — marked by the symbol W on the graph — tends to be higher on the occupational scale than that carried out by wives of workers' sons. Thirdly, it can be noted that wives of men with middle class background give up their jobs more often after marriage than wives of men with working class background.

The various illustrations presented bring to light, in my opinion several advantages and disadvantages of this method. It must be seen as a disadvantage that the graphs become over-complicated with a large number of respondents and therefore a restriction to small sample sizes and all the ensuing problems of random results must be put up with. It will therefore be necessary to safeguard individual findings by the usual quantitative tests.

One great advantage is surely the clear illustration of complex patterns which is not matched by any of the other standard methods. This happens to be true in a situation in which there are good reasons for not using a one-dimensional occupational scale for the various occupational categories. The changes of occupational improvement of such different categories as blue-collar workers, white-collar workers, civil servants, farmers and the self-employed probably obey very different rules. The various fairly arbitrarily chosen examples have shown that the graph method presents good opportunities of detecting these rules. It is therefore, at least at the stage of forming hypotheses, a useful research instrument.

In order to analyze occupational and life course patterns, one will not get around sorting the multitude of different possibilities into simplifying typologies. For this it will be necessary to perform extremely complex indexing operations and data transformations. In this respect as well the method employed should be of great help; on the one hand, because it clearly reveals the structure of the data, and on the other, because by means of plot routines the homogeneity of the types formed can easily be checked.

Finally the problem of the temporal correlation of event in different life spheres can be elegantly solved with the aid of the plot methods presented here, whereas the standard methods hardly offer any attempts at a solution.

It is surely no coincidence that just the theme „life histories“ should take up so much space at a conference of sociologists and historians. On this topic, the particular talents and abilities of the historian in analyzing singular events and developments in a multitude of relations and influencing factors, coincides too closely with the sociologist's interest in subsuming the individual case as far as possible under general structures and regularities. Therefore, the question arises as to whether the method presented here could not be utilized also in historical research with a quantitative orientation. Particularly in historical research, I can imagine problems involving an analogous data structure. For instance, in a comparative analysis of the development of different states or of subunits of individual states it could be useful to portray the available information for each individual state in its historical sequence in the form of a graph. Selected developments could then be integrated in

a multiple plot for several states. Finally, it would be possible not to plot the information according to calendar years but according to its distance in time from a decisive basic year, thus for example, according to the interval from the foundation of the state, the introduction of universal suffrage or the introduction of the first social welfare legislation. With a small number of investigation units this could certainly to a limited extent be managed by hand. If, however, the number of analysis units and of investigation problems increases, the utilization of plot routines as illustrated here could also prove to be of benefit in this context.

*Appendix 1: Scheme of the educational and occupational classification*

FB 2+	freie Berufe über 1 Mitarbeiter	self-employed 1 or more employees
FB 1	freie Berufe bis 1 Mitarbeiter	self-employed 1 employee
S 10+	Selbständige über 9 Mitarbeiter	self-employed 9 or more employees
S 2–9	Selbständige 2–9 Mitarbeiter	self employed 2–9 employees
S 1	Selbständige 1 Mitarbeiter	self-employed 1 or 0 employee
S M	mithelfende Familienangehörige	coworking family members
L 50+	Landwirtschaft über 50 ha	farmers 50 or more ha
L 20+	Landwirtschaft über 20–49 ha	farmers 20–49 ha
L 10+	Landwirtschaft über 10–19 ha	farmers 10–19 ha
L 9	Landwirtschaft bis 9 ha	farmers 9 ha
L M	mithelfende Familienangehörige	coworking family members
B HD	höhere Beamte	civil servants, high grade
B GD	gehobene Beamte	civil servants, upper grade
B MD	mittlere Beamte	civil servants, middle grade
B ED	einfache Beamte	civil servants, lower grade
A H	höhere Angestellte	top managers
A GP	gehobene Angestellte	lower managers, private sector
A GÖ	gehobene Angestellte, öffentl. D.	lower managers, public sector
A MP	mittlere Angestellte, private W.	middle white collar, private sector
A MÖ	mittlere Angestellte, öffentl. D.	middle white collar, public sector
A EP	einfache Angestellte, priv. W.	lower white collar, private sector
A EÖ	einfache Angestellte, öffentl. D.	lower white collar, public sector
WERKM	Werkmeister	supervisors, industry
ME/PO	Meister/Poliere	manual supervisors
VORAR	Vorarbeiter	foremen

F D	Facharbeiter, Dienstleistung	skilled workers, services
F G	Facharbeiter, Gewerbe	skilled workers, industry
F B	Facharbeiter, Bau	skilled workers, construction
F L	Facharbeiter, Landwirtschaft	skilled workers, farm
AA D	angelernte Arbeiter, Dienstleistung	semi-skilled workers, services
AA G	angelernte Arbeiter, Gewerbe	semi-skilled workers, industry
AA B	angelernte Arbeiter, Bau	semi-skilled workers, construction
AA L	angelernte Arbeiter, Landwirtschaft	semi-skilled workers, farm
UA D	ungelernte Arbeiter, Dienstleistung	unskilled workers, services
UA G	ungelernte Arbeiter, Gewerbe	unskilled workers, industry
UA B	ungelernte Arbeiter, Bau	unskilled workers, construction
UA L	ungelernte Arbeiter, Landwirtschaft	unskilled workers, farm
BL K	kaufmännische Lehre	commercial apprenticeship
BEANW	Beamtenanwärter	civil service preparatory courses
BL L	landwirtschaftliche Lehre	farm apprenticeship
BL G	gewerbliche Lehre	trade or industrial apprenticeship
UNI	Universität	university
ING	Ingenieurschule	engineering school
FACH	Fachschule	technical or administrative college
GYM	Gymnasium	gymnasium
REAL	Realschule	lower grade secondary school
VOLKS	Volkschule	primary school

## Computer Analysis of Life Histories from Swedish Church Records: A Case Study from the Demographic Data Base at Umeå University

There has been a steady increase in the use of collective life histories as tools for the investigation of the conditions and behavior of man in the past. It has been observed that cross-sectional data often give too little information about social change in society, stressing structures instead of processes as determinant factors.

Sometimes these changes can be traced through the study of patterns in different age groups at a given time. In other cases we need more continuous data about generations at an individual level. Usually this can be done only for a selective group in a society. The use of census-type material combined with parochial records will result in a loss of many people due to out-migration. Some researchers have tried to circumvent the problem by anticipating that on average, the out-migrants had the same characteristics as those who remained. Others have tried to make intelligent guesses about the impact of the losses, but no method has been found to correct completely for this important drawback.

From an international point of view the work which is done at the Demographic Data Base at Umeå University may have its greatest value in creating a laboratory where these alternative methods may be developed and tested. The reason for this is the design of Swedish population records from the end of the 17th century onwards. In principle the so-called church examination registers provide a continuous record for each individual with central demographic and social information.

### Sources

Parishioners were registered in Church examination registers by villages, households and families. The younger children were excluded from the earliest records, but later on in the 18th century they were also recorded. In addition to names, there is also information about the occupation of the head of the family, the dates and places of birth of all family members and the dates of marriages and deaths. There are also notes about when individuals had moved into their particular households (alone or with their families), where they came from (from another household in the

parish or from another parish), when they left the household and where they went. Absence or presence at examinations and at Holy Communion were noted annually. (Communion was often held only twice a year.) Each person who was examined received graded marks for reading from a book and for comprehension of the doctrines. Finally, any lapses in morality were also often noted, e. g., breaking any of the Ten Commandments and particularly bearing illegitimate children.

The demands from the old Swedish Central Bureau of Statistics, set up in 1749, meant that the clergy had to maintain the parish registers even more carefully. In fact, a general improvement in quality can be observed during the latter part of the 18th century. After 1800 the sources are almost always of high quality, and most of the registers have been preserved.

In addition to the examination registers, the church kept registers of births, marriages and deaths. As early as the 18th century the clergy were supposed to keep records of migrations to and from the parish. However, even if these registers ever existed on a wide scale they have mostly failed to survive. The reorganization of the Bureau of Statistics in 1860 led to the demand for more exact information about migrations which resulted in the keeping and preservation of migration registers for almost all Swedish parishes from this year onwards.

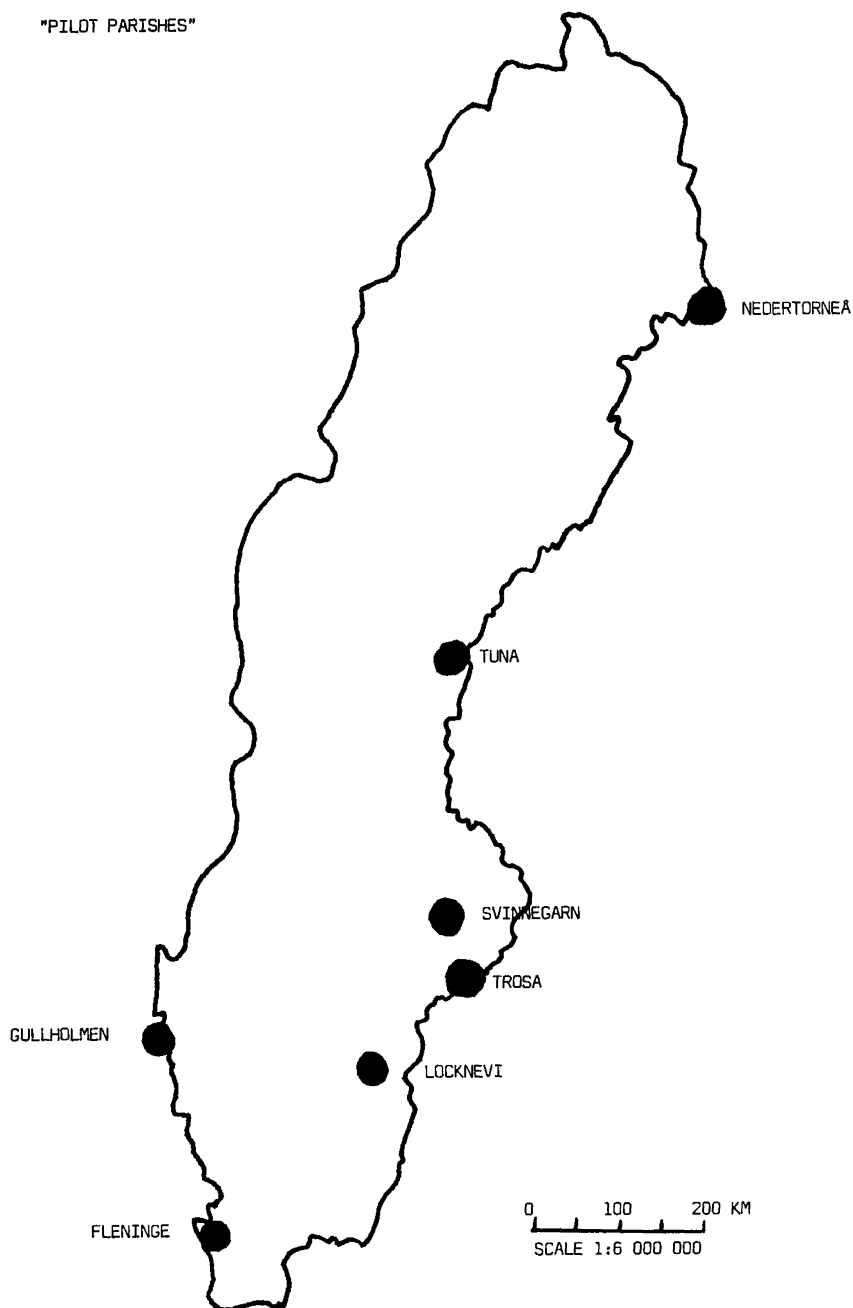
The annual registration makes it possible to trace an individual as long as he stays in a parish, and to be sure that he has been there all the time. When the source material is complete — by about 1800/1810 — it is possible to trace him from parish to parish until he dies. The possibilities for individual longitudinal studies are therefore extremely good.

## Content of the Data Base

The Data Base has, for the time being, concentrated on church records from a period between 1815 and 1895. Five parishes are now entirely at the disposal of researchers at the Data Base and two more will be available in the near future. (The seven parishes are indicated on the following map.) There are, however, some drawbacks to this strategy. For example, the subsamples of one single parish may be too small for analysis and a researcher will have to follow his migrants without computer aid. Therefore the Data Base has chosen to register several parishes forming a region around the town of Sundsvall. One of those parishes, Tuna, is on the map. Since most migrants moved within a rather small area, the majority of the migrants in the central part of the Sundsvall region may be traced by the computer. Some work may be needed to combine data for the one person who lived in various parishes, but our goal is to reduce that work as much as possible. Coding of the Sundsvall region will take another three years. In order to create a general framework for the



"PILOT PARISHES"



parishes and the region, the Data Base has also started a registration of aggregate statistics at the parish level for the seven parishes and at a somewhat broader level for the rest of Sweden (between diocese and parish). These statistics provide information about the social and demographic structure and events in five-year periods and about births, deaths and marriages each year. Death figures are given for each sex and age group separately.

No coding of other sources on either an individual or aggregate level has yet been made by the Data Base itself. In some cases independent research projects are collecting such material in cooperation with the Data Base. Two examples will be given: for one of the parishes (Locknevi) human geographers in Stockholm have created a system for the exact mapping of each house in the parish during the 19th century. Historians in the same city have been coding taxation poll data for each person who was not able to pay taxes during his or her lifetime. These data are linked on the individual level with data at the Data Base about the population of the parish of Fleninge.

As mentioned above three more years of work must be completed before the first region is fully coded. During that time different strategies must be prepared for the future. For example:

1. The Sundsvall region could be expanded further. This would make it even easier to observe the individuals as part of a larger system and at the same time make it easier to follow migrants.
2. A new region could be chosen from another part of Sweden to be coded in the same way as the first one. This would increase the possibilities for making comparisons.
3. The coding of church records could be stopped and parallel sources on the individual level could be integrated into the parishes and region already coded.
4. Migrants could be followed into their new parishes in order to create cohorts of people who had ever been within the borders of the parishes in the Data Base system.

No final choice has been made from among these priorities. There are researchers in favour of or in opposition to each of the strategies. One probable solution is that a combination of some of these approaches will be pursued.

## System of Registration and Manual Linking

All coding is done at the „production unit“ in Haparanda on preprinted cards. Each time a person appears in a source one card will be written out with the relevant data about him. 10 % of the cards are checked for errors. If the number of errors is above certain limits of acceptability the problematic variables will be corrected on all the cards from which the sample was drawn. The critical values, of course, vary depending on the character of the variable. Variables which are essential for linking are given very low critical values for error checking.

When one volume of the examination record is registered in this way the cards are sorted by hand according to given data of birth and for each person chronologically. Migration within the parish during the actual period of 5 to 10 years is represented by one card for each individual's place of residence. This operation is compared with the notes about destination of migration given on each page on which the individual has appeared. A second check is made by using information about attendance at examination and Holy Communion, so that the chronological order of the person's residence is ascertained.

A similar sorting takes place when cards from different volumes are put together, forming a chronological listing of the person's residence while in the parish. Finally, the card about birth, marriage, death and in- or out-migration are placed in chronological order within the set of cards for each person. It should be noted that normally the names are only used as a check on the other linkage variables or in the few cases when there are logical contradictions between the other variables. Cases of making one person out of two are extremely rare. Theoretically one can make two persons out of one, but since there are so many linkage variables such cases will appear very rarely. Finally, a „guide-card“ is put in front of the card file for every individual, containing a „unique identity“ of the individual, his parents and husband/wife. The unique identities are chosen from the birth register and, if this information is not available, from the first examination register. This identity is duplicated by the tape-puncher on each card belonging to the same individual.

No corrections are made of any source during this process. The manual linking gives a „proposed“ life line. Any researcher can carry out the linkage in other ways, by hand or by machine. The manual linking can be checked by the computer according to the needs of the investigator. The work on creating this system of checks has started at the research unit in Umeå. Since the work is not complete, documentation of this process has not yet published.

When the cards are sorted by hand the result is entered on tapes checked by programs in order to ensure the best possible logic, especially concerning linkage variables. Afterwards the tapes are sent to the Computer Center at Umeå University, where they are stored. The computer breaks down the material into source-oriented files. From these files one will be able to reproduce the source as it is in its primary form. Some cross-sectional studies can also be made.

## Creating Life Histories by Computer

The second step is to create new „individually-oriented files of events“ with the help of the unique identities. These files do not contain all the data within the source-files. Priorities have been made concerning selection of information which is repeated in many sources. Dates of births are, for instance, taken first from the birth register, dates of marriages from the marriage register, etc. The rules of priority may vary slightly from one period or from one parish to another, depending on the quality and accuracy of information given in different volumes. Since the priorities are made by computer they can always be documented. Should a researcher choose another system of priorities, he can do so if he is willing to pay the extra costs for programming and testing. The result is a file which is standardized for every parish. It can be used for standard programming to create statistics and listings of life histories.

The next step is to investigate the possibilities of making family life histories. Information on the „guide card“ about parents and husband/wife is checked by the computer. When the family reconstruction is accepted, new files can be created in a variety of ways according to the needs of the researcher. Genealogical searching can be done in order to describe patterns of kinship, completed family size, and age specific fertility rates can be computed, inter-generational mobility can be traced, etc.

### Information Included

The following is a list of the major information items that will be found in the so-called „individually-oriented files of events“.

- 1 Name
- 2 Date and place of birth
- 3 Parents names, dates of birth and occupations
- 4 Legitimacy of births
- 5 Migrations within the parish (years of migration and destinations)
- 6 Years of attendance at the examinations and Holy Communion
- 7 Marks for reading and comprehension
- 8 Occupational statuses
- 9 Dates of marriages and dissolution of marriage
- 10 Civil status
- 11 Dates and causes of death
- 12 In- or out-migrations

Of these data, those relating to occupation are the most difficult to attach to a specific year. A person would start as a „son“ of somebody. Later he could acquire an occupational title in a volume, but it is sometimes hard to decide which year the priest thought that his occupational career started. Changes of occupation appearing on the same page were recorded by crossing out the old occupation and writing the new one above. The frequency of this may be hard to ascertain. Occupational changes were often recorded when the priest was writing a person into a new page. The question is, however, whether the change always took place in connection with the person's migration.

Any detailed studies of the processes of intra-generational mobility will have to take these circumstances into consideration. By using other sources, e. g. taxation poll registers, the matter can often be solved in a more precise way.

Purely demographic data are generally of a very high quality and are present for the overwhelming majority of cases. Errors can therefore be detected and corrected by combining different variables and repeated information from different sources.

Years of attendance should be regarded as very accurate. The registers were kept, after all, as a result of need for this information the quality and meaning of marks are discussed by Egil Johansson in „Educational Reports, Umeå, No. 12, 1977. The History of Literacy in Sweden in Comparison with some Other Countries“.

Causes of death are of course given according to the contemporary medical knowledge of the priest. His instructions contained a list of the usual causes from which he should try to select the appropriate term for each circumstance. Generally this information is least valuable for aged persons: Their death was often classified as „because of old age“.

## File Organization

The input process has been so designed as to give a result which should contain the same set of information as the sources, even information which does not have its own data representation in the sources. When the data are organized in files, the first step therefore is to store data in so-called *source files*. These files are copies of the excerpted material as it was formed at the compilation process. The files are physically split up into five different files, one for each of the sources, but can logically be processed as one, as all the data concerning the same person have been stored using the unique identity as the key.

Since the records in the source files contain the data in the same way as it has been excerpted, all changes in forms and sources will change the contents of the records in these files. Of course this is an advantage in so far as one processes the data in order to give results based only on the information as it is represented in one of



Each person is represented by one data record containing „the most probable description of that person“. When this file is created the examination register is used as the basis for the records, which are then complemented by the other sources using the priority rule. This event file of course describes the individuals at different quality levels for different parishes depending on the completeness and the quality of the source material (and the quality of the excerpt) but the structure is the same for all parishes. The implication of this is that it will be possible to make standard programs workable for any parish, which will be of great importance in saving time for clients who want information from the Data Base. We must, however, stress the fact that the event file is only an alternative to the source files, which is why any researcher may choose input data in the way most suitable for him.

The final structure of the event file is not yet decided. When this decision is made a complete documentation will be provided. The development work is aiming to have a file containing information to satisfy the majority of Data Base clients.

## Possibilities and Drawbacks of the System

As seen from the examples above, the system can be used in different kinds of investigations. Data could be accumulated for each individual separately, for the family, the household, the village or the parish. The methodological problems concerning the first two categories are somewhat easier to solve than those of the third. In principle, each page in the examination register should contain one household. Sometimes, however, the priest had to use the page for two households, and in these cases there may be difficulties in separating the units. In some books the system did not work according to the normal standards. Servants could be entered in one row at the end of the volume without any indication of the household to which they belonged. Another problem involves deciding what the priest meant to be a household. We are not always sure that the members lived in the same house or had their meals at the same table. The older generation could have retired to a small cottage beside the main building on the farm and yet may sometimes be registered as belonging to the same household as the head of the farm.

A careful researcher therefore often has to consult other sources, mainly the taxation poll registers. In this source the unit comparable to a household was to begin with what was called a „smoke“, indicating that they cooked their meal at the same hearth. Sometimes it could, however, be used in a more legalistic sense, meaning only that the head of the „smoke“ was responsible for paying the taxes of the other members.

Because of these problems, the Data Base has chosen not to give a fixed definition of the concept „household“. Every row of the page has been given a number,

so that the researcher is able to use this information in making his own rules about the composition of households. Once he has made these rules and perhaps made completions with the help of other sources this new information could be put into a new file.

Until now the resources at the disposal of the research branch of the project have been too small. Some work on the documentation of the linking process has therefore been postponed but more will be done during the coming autumn. The manual linkage will be rested by computer.

Attempts have been made to give an algorithm which tells researchers more about the procedure. This is even more important since it has been claimed that Swedish sources offer the best opportunities of making a safe linking.

Voices have also been raised saying that the linking process is either difficult and ambiguous or simple and unambiguous. If the latter is the case, the first step in the linking could be made by computer. To date, such proposals have been rejected mainly on the grounds of the work needed from programmers. A serious test will be made in order to discover the possibilities of machine-linkage. If the advantages are greater than the drawbacks, such a linking will of course be introduced.

The great amount of information for each individual in the Swedish sources is of course valuable in many aspects, though it also creates problems. The time spent on writing cards for each individual is considerable, and this reduces the possibilities of covering a long period of a large area. Every researcher must be aware of this problem if he wants to use the Data Base. In most instances the time saved by using the machine-readable data as a starting point will be considerable, even if one has to make completions by hand.

The advantages seem, on the other hand, to be obvious: populations will be almost complete. Data are at the disposal of researchers in one single machine-readable source. The standardized process affords opportunities for checking the errors carefully. Every researcher does not have to be an expert on the primary sources. Finally, the same set of data can be used by different investigators for different purposes.

## Major Fields of Application

In order to discuss various possibilities offered by the Swedish church records and other parallel sources an international symposium was held in Umeå June 8th—11th 1977. Researchers representing different branches of science from Canada, United States, Great Britain, The Netherlands, Hungary and other countries met at the symposium. The need for the data was strongly supported by geneticists, human geographers, historians and statistical demographers.

One general impression which emerged from the symposium was that there is a



need for local studies on the individual level. Aggregate statistics were not considered to be sufficient for an understanding of the social processes which take place in history. One of the topics often mentioned was social change during the transition from an old rural to a modern industrialized and urbanized society. The impact of environment upon man and man's attempts to adjust his environment according to his needs are sometimes more easily discerned on a local level. This is one of the major fields of research for many human geographers and ethnologists.

Human migration, family building, occupational mobility, the development of popular education, among other things, seem to be hard to explain without individual longitudinal studies. Certainly, the reader can find a lot of other questions which are as important to answer as those mentioned, where the life histories of individuals can provide one piece of the puzzle. The Data Base at Umeå University was designed from the beginning to be a systematic project at the disposal of every researcher irrespective of his origin. Sometimes it can be used as a laboratory where methods are tested and comparisons made with findings in other investigations. Every scientist is therefore welcome to request our services. There are some minor restrictions concerning the use of data because of the Swedish law about individual privacy, but these have no important bearing on 19th century material.

The customer will pay for the time needed by the programmer to design the file according to his special needs and for the time consumed by the computer. Contact should be made as soon as possible, even if no promises can be given yet about the time of delivery. Since this report has been written for an international audience, it is important to stress the weight that is given to the attitudes of foreign researcher in the long-term planning of the Data Base.

## Documentation

The Demographic Data Base: A Short Introduction.

The Demographic Data Base: Sources and Codes.

Further Documentation is under preparation describing the different error checks, the process of linkage and the file-system.

This documentation can be ordered from:

The Demographic Data Base  
Management and Research Unit  
Humanisthuset  
University of Umeå  
S-901 87 UMEÅ Sweden

Anyone wishing to be on the mailing-list of the Data Base is welcome to apply to the same address.

Quantitative Analyses of Collective Life Histories:  
The Case of the Social Democratic Candidates for the  
German Reichstag 1898–1912

The inquiry into the Social Democratic Reichstag candidates is a segment of a comprehensive project on the collective life-histories of the functionaries of the social democratic working-class movement in Germany from 1890 to 1914<sup>1</sup>.

Altogether four inquiries are planned:

the Social Democratic Reichstag candidates 1898 to 1912

the Social Democratic party convention delegates 1897 to 1913

the delegates to the general union congresses 1892 to 1914

the salaried personnel of the party and of the unions in the years 1914 and 1927.

The total population comprises more than 15 000 functionaries. The collection of data is on-going; the numerous problems of my investigations which principally arose in the reconstruction of historical biographies should not be discussed here. My presentation deals with two sets of problems:

- the analysis of the occupations and
- analytical approaches, particularly the comparison of „longitudinal analysis“ with „cross-sectional analysis“ and „individual analysis“ with „aggregate analysis“.

<sup>1</sup> This paper was presented at the QUANTUM-SSHA-Conference „Quantification and Methods in Social Science Research: Possibilities and Problems with the Use of Historical and Process-Produced Data“, held in Cologne (W.Germany), 10–12th Aug. 1977. The conception of this presentation dealt only with some particular techniques and methods and does not feature any special results of my analyses; for a more detailed discussion of my results see: Schröder, Wilhelm Heinz, *Die Sozialstruktur der sozialdemokratischen Reichstagskandidaten 1898–1912*, in: *Herkunft und Mandat, Beiträge zur Führungsproblematik in der Arbeiterbewegung*, Frankfurt a. M./Köln 1976, pp. 72–96; Schröder, W. H., *Probleme und Methoden der quantitativen Analyse von kollektiven Biographien*, in: Best, Heinrich, and Mann, Reinhard (eds.), *Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung*, Stuttgart 1977, pp. 88–125; Schröder, W. H., *Arbeitergeschichte und Arbeiterbewegung, Industriearbeit und Organisationsverhalten im 19. und frühen 20. Jahrhundert*, Frankfurt a. M./New York 1978. — This paper was translated by Robert Dees (Washington D. C.).

## Analysis of Occupations

Biographical data require meticulous and critical preparation of sources. In so far as official documents are available — such as family and civil status registers — the data appears to be relatively reliable. But other official documents — like trial records, arrest warrants, personal files — contain on the other hand mostly a considerable mass of misinformation. The evaluation of obituaries as they were customarily published in the Social Democratic daily newspapers appears particularly problematic. Such obituaries which most importantly present a major biographical source for the lesser known candidates, are, however, usually sketchy with regard to precise, usable data since they emphasized primarily the general, the paradigmatic, while the individual life story was treated as of secondary importance. A similar intent can be ascertained in numerous autobiographical contributions by class-conscious workers of this period<sup>2</sup>.

The information given by the relevant functionaries themselves emphasized functional character — for example in parliamentary or intra-party manuals, in *curricula vitae* or in official address-books. Depending on the type of publication, certain biographical data would consciously be partly or wholly omitted, modified or distorted. One of the most vivid examples of this is the manipulation of details concerning occupation: often the point in time in which the person in question exercised the stated occupation and particularly the position occupied remain unclear.

The functional character of the biographical details diminished the value of the information and the quantitative basis of the attributes thereby affected. However, it allows for the elimination of contradictions and gaps within the data through comparison with and supplementation through data from other sources. Beside the narrower occupation analysis this operational imprecision impairs principally the results of the studies on inter- and intragenerational mobility<sup>3</sup>. On the one hand the simplifying identification of the father's profession or occupation as an indicator of social origin is not always completely convincing. On the other hand the inexactitude as well as the deficiency of the transmitted information limits the researcher to mere assumptions about origin and intragenerational mobility. First the complementary inclusion of further attributes like school education and occupational training can specify the statements about social origin.

<sup>2</sup> Cf. the autobiographical anthologies: Emmerich, Wolfgang (ed.), *Proletarische Lebensläufe, Autobiographische Dokumente zur Entstehung der Zweiten Kultur in Deutschland*, 2 vols., Reinbek 1974 and 1975; Klucsarits, Richard, and Kürbisch, Friedrich G. (eds.), *Arbeiterinnen kämpfen für ihr Recht, Autobiographische Texte rechtloser und entrechteter „Frauenspersonen“ in Deutschland, Österreich und der Schweiz des 19. und 20. Jahrhunderts*, Wuppertal 1975.

<sup>3</sup> For a detailed bibliography on social mobility, social stratification, social change in history see: Wehler, Hans-Ulrich (ed.), *Bibliographie zur modernen Sozialgeschichte*, Göttingen 1976, pp. 96–164 and 241–261.

Take the example, well-known in Germany, of the career of the Bohemian worker Wenzel Holek, who in the course of his life changed positions and occupations frequently and even carried on several jobs at the same time<sup>4</sup>. It will be understood that this continuous occupational change is not functionally suited to a fixed format which sufficiently describes all variants. A multiple occupation can only be represented by reducing it to the principal occupation.

Notwithstanding the definitional difficulties of understanding what was „Arbeit“ (work) and „Beruf“ (profession/occupation)<sup>5</sup>, a reduction to principal occupational positions within a career is necessary to the investigation of generational mobility. In the case of the Reichstag candidates this, for the most part, means the determination of the „trained“ occupation (or „first“ occupation), of the (at the time of the candidature) „exercised“ occupation (or „actual“ occupation) and among the salaried personnel of the working-class movement the determination of the last provable occupational position before the first employment as an employee in the working-class movement.

The indicated occupational positions in the occupation-statistics have to be combined and classified in a second step in order, on the one hand, to condense the information content of the acquired detailed results, and, on the other hand, to avoid the mere description of results. To establish comparability with the other populations, it is useful to compare the occupational statistics of the Reichstag candidates with the membership statistics of the socialist union movement and with the official occupation and trade statistics.

Without being able to go thoroughly into the general problems of cross-level analysis<sup>6</sup>, a comparison of the Reichstag candidates can be questioned in terms of the following objections (see Tables 1, 2 and 3)<sup>7</sup>:

1. Even if the „trained“ or „exercised“ occupation can unequivocally be determined, the assignment of this occupation to the appropriate union is not always realizable. The defined occupational recruiting fields of the individual unions

<sup>4</sup> Holek, Wenzel (1864–1935), *Lebensgang eines deutsch-tschechischen Handarbeiters*, ed. by Paul Göhre, Leipzig/Jena 1909; Holek, Wenzel, *Vom Handarbeiter zum Jugenderzieher*, ed. by Theodor Greyerz, Jena 1921.

<sup>5</sup> Cf. e. g.: Daheim, Hansjürgen, *Der Beruf in der modernen Gesellschaft, Versuch einer soziologischen Theorie des beruflichen Handelns*, 2nd ed., Köln/Berlin 1970; Neuloh, Otto, *Arbeits- und Berufssoziologie*, Berlin/New York 1973.

<sup>6</sup> Cf. e. g.: Hummell, Hans J., *Probleme der Mehrebenenanalyse*, Stuttgart 1972; Scheuch, Erwin K., *Entwicklungsrichtungen bei der Analyse sozialwissenschaftlicher Daten*, in: König, René (ed.), *Handbuch der empirischen Sozialforschung*, 3rd. ed., Stuttgart 1973, Vol. 1, pp. 209–219; Pappi, Franz Urban, *Aggregatdatenanalyse*, in: Van Koolwijk, Jürgen, and Wicken-Mayser, Maria (eds.), *Techniken der empirischen Sozialforschung*, Vol. 7, München/Wien 1977, pp. 78–110.

<sup>7</sup> The membership statistics of the German Social Democratic unions (Table 1) from: Ritter, Gerhard A. and Tenfelde, Klaus, *Der Durchbruch der Freien Gewerkschaften zur Massenbewegung im letzten Viertel des 19. Jahrhunderts*, in: Ritter, Gerhard A., *Arbeiterbewegung, Parteien, Parlamentarismus*, Göttingen 1976, pp. 99–100.

Table 1: Cross-Level Comparison of Occupations

1	2	3	4	5	6	7
1	Metal Workers	101	15.0	14.9	12.3	4.6
2	Woodworkers	98	14.5	12.2	5.9	7.7
3	Shoemakers	48	7.1	3.5	2.2	6.1
4	Printers	46	6.8	7.5	1.1	26.8
5	Building Workers	40	5.9	6.7	8.9	2.8
6	Tobacco Workers	30	4.5	5.8	1.8	12.0
7	Tailors	23	3.4	3.1	3.5	3.4
8	Textile Workers	20	3.0	6.8	10.7	2.4
8	Sales Workers	20	3.0	0.3	6.6	0.2
10	Miners	16	2.4	2.0	7.4	1.0
11	Painters	15	2.2	2.7	1.1	9.3
12	Transport Workers	13	1.9	2.6	1.7	5.8
13	Carpenters	12	1.8	3.6	1.4	9.7
14	Sculptors	10	1.5	1.2	?	?
14	Bookbinders	10	1.5	1.5	0.7	7.6
14	Operatives, Work. (Gen.)	10	1.5	2.6	8.5	1.2
17	Brewers, Millers	9	1.3	2.9	2.2	4.9
17	Saddlers	9	1.3	0.6	0.5	4.7
19	Lithographers	8	1.2	1.6	0.4	15.5
20	Paperhangers, Decorators	7	1.0	0.3	0.2	5.3
21	Bakers, Confectioners	5	0.7	0.6	2.1	1.1
21	Glass Workers	5	0.7	0.9	0.8	4.5
21	Glovers, Leather Workers	5	0.7	2.6	0.8	12.4
24	Clerical Workers	4	0.6	0.1	?	?
24	Potters	4	0.6	1.4	0.5	10.0
24	Farm Workers	4	0.6	—	—	—
27	Butchers	2	0.3	?	?	?
27	Hairdressers, Barbers	2	0.3	0.3	0.4	2.4
27	Gardeners	2	0.3	0.1	—	0.7
27	Porcelain Workers	2	0.3	2.8	0.6	15.8
31	Coopers	1	0.1	1.6	0.3	20.9
31	Slaters, Tilers	1	0.1	0.6	0.3	8.5
31	Waitors	1	0.1	0.7	3.9	0.7
31	Hatters	1	0.1	1.1	0.3	12.6
31	Quarrying Workers	1	0.1	1.8	1.6	4.2
31	Musicians	1	0.1	—	?	?
—	Academic/Bourgeois Occ.	62	9.2	—	—	—
—	Unknown	24	3.9	—	—	—
Total		673	100.0%			

1 = Candidates: No. Rank Occupation (Freq. Distribut.)

2 = Candidates: assigned Unions/Occupations (Offic.): Name of Occupation

3 = Candidates: Absol. Freq. 1898–1912

4 = Candidates: Relat. Freq. 1898–1912

- 5 = Unions: Union's Percentage 1895 (All Unions = 100) (n = 255 521)  
 6 = Official Statistics: Occupation's Percentage 1895 (All Occupations = 100) (n = 6 810 666)  
 7 = Unions/Official Statistics: Percentage of organized workers within the recruiting field of the union 1895

Classification: 48 German craft, industrial, and general unions in 1912

*Table 2: Frequency Distribution of the „First“ Occupation 1898–1912*  
 (Classification: Industrial Branches)

Industrial Branch	Election 1898	Election 1903	Election 1907	Election 1912
Metal	37 13.2 %	42 13.1 %	55 16.4 %	59 16.1 %
Building	13 4.6 %	31 9.7 %	29 8.6 %	45 12.3 %
Textile	8 2.9 %	9 2.8 %	9 2.7 %	6 1.6 %
Trade/Transport	11 3.9 %	14 4.4 %	16 4.8 %	18 4.9 %
Mining	1 0.4 %	4 1.2 %	7 2.1 %	5 1.4 %
Clothing/Shoes/Leather	38 13.6 %	35 10.9 %	39 11.6 %	41 11.2 %
Woodworking/Furnishing	50 17.9 %	59 18.4 %	55 16.4 %	61 16.7 %
Food/Drinking/Tobacco	25 8.9 %	24 7.5 %	23 6.8 %	23 6.3 %
Quarrying/Glass/Pottery	4 1.4 %	4 1.2 %	7 2.1 %	6 1.6 %
Printing/Paper	30 10.7 %	30 9.3 %	27 8.0 %	27 7.4 %
Other Occupations	63 22.5 %	69 21.5 %	69 20.5 %	75 20.5 %
Total	280 100.0 %	321 100.0 %	336 100.0 %	366 100.0 %

Election 1898: 280 socialist candidates for 395 areas

Election 1903: 321 socialist candidates for 395 areas

Election 1907: 336 socialist candidates for 397 areas

Election 1912: 366 socialist candidates for 397 areas

**Table 3: Frequency Distribution of the „Actual“ Occupation 1898–1912**  
(Classification: Adapted Stratification)

Stratum	Election 1898	Election 1903	Election 1907	Election 1912
Unskilled/Semiskilled Workers	2 0.7 %	3 0.9 %	0 0.0 %	0 0.0 %
Skilled Workers/Sales Craftsmen/ Clerical Workers	26 9.3 %	31 9.7 %	16 4.7 %	6 1.6 %
Working-Class Movement Employees	148 52.9 %	203 63.2 %	249 73.9 %	310 84.7 %
Trade Unions	24 8.6 %	39 12.1 %	68 20.2 %	99 27.1 %
„Secretariat of Workers“	7 2.5 %	21 6.6 %	29 8.6 %	34 9.3 %
„Party Secretariat“	3 1.1 %	6 1.9 %	23 6.9 %	40 10.9 %
Party Editors/Writers	63 22.5 %	71 22.1 %	57 16.9 %	67 18.3 %
Party Employees	33 11.8 %	39 12.1 %	44 13.0 %	39 10.7 %
Co-operative Employees	7 2.5 %	18 5.6 %	17 5.1 %	17 4.5 %
Health Insurance Employees	11 4.0 %	9 2.8 %	11 3.2 %	14 3.9 %
Little Tradesmen/Producers/ Middle Officials	91 32.5 %	72 22.4 %	57 16.9 %	35 9.6 %
Farmers	1 0.3 %	0 0.0 %	0 0.0 %	0 0.0 %
Academic/Bourgeois Occupations	10 3.6 %	10 3.1 %	14 4.2 %	15 4.1 %
Unknown	2 0.7 %	2 0.6 %	0 0.0 %	0 0.0 %
Total	280 100.0 %	321 100.0 %	336 100.0 %	366 100.0 %

often overlapped despite agreements with each other over suitable boundaries — e. g. the fields of an intrasectoral industrial union vs. a multisectoral occupational or branch union<sup>8</sup>.

2. An individual could work successively in various industrial sectors and thereby within different union recruiting fields even without occupational change simply by changing the place of work. This was particularly common among unskilled workers, who could work in many different industrial sectors and thereby also be organized, into individual unions. But this applies just as well for a skilled woodworker, who actually should have been assigned to the wood workers union, but in fact — as the corresponding membership statistics make plain — was employed in the metal industry and consequently was organized in the German metal workers union<sup>9</sup>.
3. Without concrete information on actual organizational membership, the numerous workers who changed occupations, cannot be clearly assigned to a specific occupation category. These cases hypothetically will be assigned to the union to which they probably belonged on the basis of their occupational training. This applies all the more to the numerous self-employed and salaried personnel of the working-class movement, who as a rule maintained their membership in the union of their former occupation. Unions with multi-sectoral recruiting fields — as for example the factory workers union (General Union of Operatives) — would through this method of classification be badly underrepresented statistically. Only when the membership in such an union is indicated by the life-histories of the candidates does a corresponding assignment follow. An unskilled textile worker, for example, ordinarily is assigned to the textile and not to the factory workers union<sup>10</sup>.
4. The difficulties of comparison of the membership statistics of the unions with the official statistics of the German census of occupations in regard to the degree of organization of union recruiting fields was recently discussed<sup>11</sup>. Similar concerns also apply to the use of indications of occupation as simplified social stratum and class criteria (see Table 3). The necessity of compiling comparable statistics, as a rule, leads to the use of a static social stratum or class model and to

<sup>8</sup> Already the first trade-unions-congress (1868) of the socialist „Deutscher Arbeiterschaftsverband“ tried to define such boundaries; but nearly every following congress had to discuss new corrections of the boundaries; cf.: Barthel, Paul, Handbuch der deutschen Gewerkschaftskongresse, Dresden 1916, pp. 192–217.

<sup>9</sup> In 1913 e. g. there was one percent of the members of the German metal workers union (1913: 0.55 Mill. members) non-metal-workers, cf.: Jahrbuch des Deutschen Metallarbeiterverbands 1913, Stuttgart 1914, p. 43.

<sup>10</sup> For example August Brey (1864–1940), the leader of the general union for operatives (unskilled and semiskilled workers), worked first as a skilled shoemaker (his „trained“ occupation) within the „Handwerk“, later he changed and worked as an operative within the industry (his occupational position before the first employment as an employee of the union).

<sup>11</sup> Cf.: Ritter, Durchbruch, pp. 99–100.



the neglect of the historical dimension, which actually requires a dynamic model for the adequate depiction of historical change.

Taking into account the father's occupation and the early birth dates of the oldest Reichstag candidates, the time period of my investigation of the candidate analysis encompasses more than a century. In the course of this century, however, social and economic functions, ascribed status etc. of the traditional professions/occupations changed or ceased to exist, while numerous new occupations/professions came into being and restructured the traditional social strata. This change is documented most clearly in the traditional separation of the professions/occupations along the lines of the legal status of the self-employed. These categories increasingly lost their meaning as relatively clear status criteria at the latest with the social and economic change of the German „Handwerk“ and trade and with the general introduction of freedom of choice of occupation, and lived from then on primarily as elements of ideology.

However, in the case of the working-class movement the legal right of self-employment was of special social and political importance in the early period of the movement. This right of self-employment permitted those workers who had been fired or boycotted for political reasons to continue promoting the activities of the working-class movement<sup>12</sup>. Nevertheless, there is no patent recipe for the crux of the cross-level analysis problems; the researcher is limited to adapting the stratum criteria of his own investigation to the fixed classification of comparative historical statistical publications. The necessity of having to compare also led the official statistics of the German Reich to only rarely deviate from the method of survey practiced since 1875 and 1882. Thereby the German Reich statistics steadily lost its capability to adapt to and validly measure historical change using dynamic categories of the observable phenomena.

## Methodology

The theoretical approaches (e. g. organization theory, work and conflict theory) followed in the framework of the whole project cannot be presented here<sup>13</sup>; but the delimitation of the individual fields of inquiries according to different groups of functionaries allows at least the positional approach of historical elite-research to be recognized<sup>14</sup>. Without taking up the controversial discussion over the content of

<sup>12</sup> Cf.: Schröder, Sozialstruktur, pp. 85–88.

<sup>13</sup> Cf.: Schröder, Arbeitergeschichte.

<sup>14</sup> For the elites-studies cf. e. g.: Eldersveld, Samuel J., Political Parties, A Behavioral Analysis, Chicago 1964; Valen, Henry, and Katz, Daniel, Political Parties in Norway, A Community

the concept „elite“, the classification of most partial investigations in the framework of empirical research of elites – for lack of a correspondingly defined scientific field – appears justified. The investigated functionary groups in the figurative sense – taken as a whole – can be described as „functional elites“ within the working-class movement. In keeping with this methodological positioning, questions are posed as to the social profile, the selection and circulation as well as the power structure of the investigated functionary groups. Thereby precisely in the realm of the working-class movement the question of the elites as „index of the social structure“<sup>15</sup> acquires a central meaning.

The disadvantages of this method of positional analysis are sufficiently known. They arise for the most part through the simplifying assumption that formal and actual power structures are identical and that the formal authority to which a position will be attributed is identical to the actual exercise of power of the position holder. While the contemporary research of elites at least can serve as supplementary and corrective of the other two methods (decision-making approach and reputational approach), the historian remains limited mostly to the positional approach merely for lack of sources. The observation of historical decision-making processes and actual power structures is possible only with great difficulties, and the interviewing of historical participants is rarely possible.

In this sense the more narrow selection process, for example, which finally led to the nomination of the Reichstag candidates can be reconstructed only in cases in which the conflicts were so spectacular as to break out into public view<sup>16</sup>. Otherwise the accounts of the actual nomination process moved on an ideal-typical level or the accounts are only hypothetical in nature as a mediate result of the analysis.

An historical analysis of the personnel change and/or of the structural change of the candidate group cannot be realized in a dynamic approach. That is, there exists no possibility of following the historical development continuously over a period of time. A certain number (at least two) chronologically successive static points in time must be produced and compared with each other. In such a methodology im-

Study, Oslo/London 1964; Zapf, Wolfgang, *Wandlungen der deutschen Elite, Ein Zirkulationsmodell deutscher Führungsgruppen 1919–1961*, München 1965; Zapf, Wolfgang (ed.), *Beiträge zur Analyse der deutschen Oberschicht*, München 1965; Scheuch, Erwin K., *Führungsgruppen und Demokratie in Deutschland*, in: *Die Neue Gesellschaft*, 13. Jg. (1966), pp. 356 ff.; Ammon, A., *Eliten und Entscheidungen in Stadtgemeinden, Die amerikanische „Community-Power“-Forschung und das Problem ihrer Rezeption in Deutschland*, Berlin 1967; Barnes, Samuel S., *Party Democracy. Politics in an Italian Socialist Federation*, New Haven/London 1967; Ludz, Peter Ch., *Parteilite im Wandel, Funktionsaufbau, Sozialstruktur und Ideologie der SED-Partei-führung*, 3rd. ed., Köln/Opladen 1970; Zoll, Rainer (ed.), *Gemeinde als Alibi*, München 1972; Witjes, Claus, *Gewerkschaftliche Führungsgruppen*, Berlin 1976.

<sup>15</sup> Cf.: Zapf, *Wandlungen*, p. 207.

<sup>16</sup> Cf. e.g. the case of Paul Göhre (member of the German „Reichstag“: 1903, 1910–1918): Heine, Wolfgang, *Demokratische Randbemerkungen zum Fall Paul Göhre*, in: *Sozialistische Monatshefte*, 8 (1904), pp. 281–291.

precisions must be tolerated, since short term change in the unobserved periods in-between cannot be identified, such that the comparison of the net movement possibly can result in ungrounded or imprecise explanations of structural change or of stability. The closer to each other those momentary points of analysis are, the more exact and differentiated can such structural processes of change be traced on the basis of the construction of operational factors. If the points of survey in time are sufficiently close and regular, than ultimately time-series as well as longitudinal analyses is possible<sup>17</sup>.

The delimitation of the positional elite „Reichstag candidates“ to be investigated anticipates a series of methodological decisions:

1. The Reichstag elections took place in the rule every four or five years (1898, 1903, 1907, 1912). The positional elite „Reichstag candidates“ remains therefore not continuously determinable, rather constituted itself exclusively for the electoral campaigns. The „lifetime“ of the group was limited to the few months or weeks before the election; the „group-lifetime“ therefore does not describe the variable „time“ as a continuum, but rather as a regularly repeated period of time. Ultimately, the impossibility of surveying continuous data even in this limited time period requires the compression of the „group-lifetime“ to the point in time of the election, resulting in the decision for cross-sectional analysis.
2. Four elections accordingly mean four points of survey and four cross-sectional analyses. In order to be able to ascertain the change within the structure of the Reichstag candidates group, these four cross-sectional analyses must be compared with each other.

The first methodological deficiency is that an important preliminary decision for the procedure of cross-sectional analysis cannot be problematized at all: the determination of the „optimal“ spaces between the points in time of the cross-sections. At least for all analytical questions which go beyond the more narrow election analysis, one must carefully check to what degree the proposed analytical cross section on the time axis can validly measure the change. The second methodological deficiency is that the population for the inquiry does not remain constant over the course of the four elections but changes in each election by relatively high rates. Not only do these transformations create numerous difficulties in analytical comparison, but also these changes in the population complicate statements on the rank of the position criterion „Reichstag candidates“. For example, does the overrating of the electoral success within the Social Democratic Party find its equivalent in an equally high evaluation of the Reichstag candidacy and their personal representatives? Did candidacy represent a high reputational status that was recognized by the party? Did candidacy presuppose a certain career in the party or in the unions? To what degree did candidacy have influence later on the professional career as a salaried functionary in the organiza-

<sup>17</sup> For the general problems of time-series and longitudinal analyses see e. g.: Dierkes, Reinolf, *Die Analyse von Zeitreihen*, in: Van Koolwijk, Jürgen, and Wicken-Mayer, Maria (eds.), *Techniken*, pp. 111–169.

tions of the working-class movement? What happened to those candidates who were not renominated? In short: the inquiry of the life-histories of those candidates who were no longer or were not yet nominated is just as important as the investigation of those who were either renominated or were nominated for the first time.

This preliminary decision for carrying out cross-sectional analysis will also be favored through the use of SPSS<sup>18</sup>, which is tailored to the comparison of cross-sectional analysis in the above mentioned sense. This preliminary decision simultaneously means deciding against an analysis of the individual biographies evaluated within their own context, and in favor of the dissection of the biographies, and then evaluation of the aggregated mass of single data independent of their original context. The goal of my inquiry is to work out a collective biography of the group of Reichstag candidates. The subject of investigation of this collective biography consists of the sum of the individual life-histories.

The first procedure of comparatively evaluating these individual life-histories consists of individual analysis which at least through the SPSS performance is not explicitly provided for. The individual analysis examines each individual life-history separately for recognizable individual correlations and courses of events and compares those individual results with the other life-histories. Not until the end are these individual results eventually aggregated. The second procedure consists of the aggregate analysis<sup>19</sup> which is facilitated by SPSS. The aggregate analysis breaks up the individual connections and separates the life histories into individual data which are then compares and/or aggregated with the individual data of other life-histories. The criticism which is often made up by historians against the aggregate analysis, that man did not exist as a group or as a carrier of aggregated single data, but rather as a whole individual, can be rebutted precisely with the aid of individual analysis.

For analysis of collective life histories there exists therefore, a pressing need to write a program package which fulfills the differentiated requirements of biographical individual analysis, particularly:

1. Only one complete coding of the data onto cards, not a separate coding for every cross-sectional analysis;
2. Freefield format with variable record-lengths which is reconciled as close as possible to the usual form of biographical presentation;
3. Structuring into different fields of data and accounts dimensioned by time which allows the optional cross-sectional-analysis for all variables dependent on time;
4. Possibilities of comparison of individual sequences and connections.

<sup>18</sup> Cf.: Nie, Norman H., et al., SPSS, Statistical Package for the Social Sciences, 2nd ed., New York 1975; Beutel, Peter, et al., SPSS, Statistik-Programm für die Sozialwissenschaften (Kurzbeschreibung Programmversion 6), Stuttgart/New York 1976.

<sup>19</sup> For a discussion within a general context see e. g.: Pappi, Franz Urban, Aggregatdatenanalyse, pp. 78–110.

## Configuration-Sequence-Analysis (CSA)

Of course, the research cannot wait for the development of such a program package. In the following I offer no replacement for such a package, but describe several assisting operations on the basis of SPSS. The conscious limitation to the discussion of possibilities of SPSS reflects not only the experiences that I have personally had with SPSS, but rather the comprehensive availability of SPSS. This program has been implemented in almost all the German university computer centers and is used by the majority of historians in Germany<sup>20</sup>.

The starting point of my reflections is the so called „configuration-frequency-analysis“ (CFA), as was developed by J. Krauth and G. A. Lienert and which is available in a few computer centers in Germany<sup>21</sup>. The CFA certainly presents a considerable improvement over existing multivariate analysis procedures. In contrast to the other procedures (like factor-, path-, cluster-analysis and multiple regression) the CFA presents a parameter-free procedure and presupposes no distribution forms whatsoever. Without wanting to go any deeper here into the various performances of the CFA, the CFA permits the construction of attributes-configurations; that is, the combination of various classes of determined attributes into a configuration, by which assessments of certain configuration frequencies will be made easier through the printing of expected probabilities of those configurations. At least two reasons support the adaption of important elements of the CFA by SPSS-users:

1. Historical data are surveyed to a large extent only at nominal-scale level, and therefore are not utilizable with the customary multivariate analysis procedures which presuppose a higher scale-level<sup>22</sup>. This caveat applies particularly to data from life-histories.

<sup>20</sup> For the distribution of SPSS in W. Germany see: Beutel, Peter, et al., SPSS, p. 163.

<sup>21</sup> See: Krauth, Joachim, and Lienert, Gustav Adolf, KFA, Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin, Freiburg/München 1973; Roeder, Burkhard, Die Konfigurationsfrequenzanalyse nach Krauth und Lienert, in: KZfSS, 26(1974), pp.819–844; an applied example for CFA with historical data: Best, Heinrich, Die quantitative Analyse inhaltlicher und kontextueller Merkmale historischer Dokumente, in: Best, Quantitative Methoden, pp. 162–205.

<sup>22</sup> For a theoretical and empirical comparison of two recently suggested models (HARDER's DO and GOODMAN's ECTA) and the development of both models as an new approach see: Küchler, Manfred, Multivariate analyzing of nominal-scaled data, in: ZfS 5(1976), pp. 237–255.

<sup>23</sup> This phobia represents one of the most important problems for teaching quantitative methods on University. Up to now the didactics of quantification has been rather neglected; as a first step for solving this problem the QUANTUM-Action-Group „Didactics of Quantification“ (DidQuant) has been established, cf.: QUANTUM Information, No. 2 (May 1977), p. 11, and 4 (Oct. 1977), p. 4.

2. A simplified CFA with assistance from SPSS leads not only to a spectacular demonstration effect of multivariate connections through the emergence of certain types of configurations, but also helps to reduce the still ineradicable statistics-phobia of German history students<sup>23</sup> and to bring about recognition of the necessity of such procedures.

The CFA according to Krauth and Lienert, however, starts from a static model and still does not signify any decisive progress for the analysis of life-histories. Only by addition of elements from the CFA and from time-series analysis can we make any progress. The CFA presupposes the following classification standards for the codes of the variables:

1. *Unambiguousness* — the code of a variable can be attributed to each object
2. *Exclusivity* — only one single code of a variable applies for each object
3. *Completeness* — each object must own a code of the variable

However, the postulate of completeness can rarely be fulfilled with historical data, so one would do better to speak of a typologization rather than a classification. In addition there are:

4. *Chronology* — the positions of the configuration must be chronologically ordered.
5. *Thematic Homogeneity* — the position series must fulfill the conditions that the position sequence only follows the development of *one* phenomenon over time.

To recapitulate, the new configuration-sequence-analysis (CSA) can be defined as a „time-ordered sequence of configuration positions for the observation of a single phenomenon in the course of time“<sup>24</sup>.

I have chosen an example from bureaucracy research from my area of investigation to demonstrate the simplified CFA or the CSA. The prevalent stereotype of the research on the history of the German working-class movement is that the Social Democratic movement organizationally consolidated itself between 1890 and 1914 and developed an overgrown bureaucracy<sup>25</sup>. Accordingly one traces numerous developments within the workers movement back to this bureaucratization; the catchword „bureaucratization“ serves generally as an unreflected explanation for negative developments.

Taking Max Weber's bureaucracy model as a basis for our analysis, „bureaucracy“ can be nailed down and measured by means of 11 main characteristics<sup>26</sup>. Of these

<sup>24</sup> For definitions of „time-series“ see: Dierkes, Analyse, p. 111.

<sup>25</sup> Cf. the minimum-number of working-class-employees:

party/trade unions: 659 (1903), 1871 (1907) and 3297 (1912)

co-operative soc.: 7081 (1903), 12783 (1907) and 26402 (1912)

for the complete statistics see: Fricke, Dieter, Die deutsche Arbeiterbewegung 1869–1914, Berlin (DDR) 1976, pp. 741, 776.

<sup>26</sup> Cf.: Weber, Max, Wirtschaft und Gesellschaft, 5th ed., Tübingen 1972, pp. 120–130, 541–579 and 825–837; for a discussion on Weber's approaches cf. e.g.: Mayntz, Renate, Max Webers Idealtypus der Bürokratie und die Organisationssoziologie, in: Mayntz, Renate (ed.), Bürokratische Organisation, Köln/Berlin 1971, pp. 27–35.

characteristics, however, at best only half can be found in the life-histories of the holders of bureaucratic positions. Next to the main occupational activity and the fixed pay for the exercise of official functions stands one of the most important attribute of bureaucratization — according to Max Weber — the formation of fixed sequential career patterns with institutionalized steps of higher and lower ordered bureaucratic positions.

To see how I have examined the characteristic „career patterns“ for my population, please inspect Tables 3 and 4 and Figures 1, 2, 3 and 4. The prevalent methodology for cross-sectional analysis requires, that the attributes of each case from the historical time perspective of every cross-sectional analysis must be examined for its specific form and for its sequence in time. The variable „Position Union Employee“ would be answerable, with regard to the life-history of the candidate, either in the form of a dummy-variable (yes/no) or ordered according to time (e. g. „earlier“, „at the time of candidacy“, „later“). However, this says very little about the sequence of the positions of union employees. The question of kind, form and course of career patterns remains unanswered.

The various career stations of the variable „sequence“ will be punched in consecutively with the respective particulars of the occupied position and the year of the beginning and end of tenure. Thereby a variable record-length is possible. Particularly with the help of the SPSS-Compute-Card the whole position-codes can be combined into the position-sequence-variable which can have a maximum of 12 position-columns. The sequences can be clearly documented through a simple frequencies-run. However, indication of the probabilities to expect — as with CFA — is lacking here. Nevertheless, the possibility exists in principle to determine each year of the time-period as a section of a cross-sectional analysis and not only the year of the elections. In this manner one can produce time-series for particular attributes by juxtaposition of periodical cross-sectional analyses. All variables which can be preserved in a time span are suitable for such configurations. Configuration-combinations with static individual attributes — like place of birth (see Figure 4) — can be realized at will.

My results on the development of career patterns within the working-class movement organizations, however, can be summarized as follows. Contrary to the prevalent research stereotype, only the beginnings of a bureaucratic development are to be found, a fixed and differentiated career pattern did not exist. Not even the emergence of the „party secretaries“ as a clear goal-position constitutes a career pattern. The results of these analyses on bureaucracy perhaps will help to revise Robert Michels' well-known bias concerning the bureaucracy and oligarchy within the Social Democratic working-class movement in Germany before World War I<sup>27</sup>.

<sup>27</sup> Cf.: Michels, Robert, *Zur Soziologie des Parteiwesens in der modernen Demokratie*, 1st ed., Leipzig 1911 (reprint of the 2nd ed., Stuttgart 1957); Michels, Robert, *Die deutsche Sozialdemokratie, Parteimitgliedschaft und soziale Zusammensetzung*, in: *Archiv für Sozialwissenschaft und Sozialpolitik* 23 (1906), pp. 471–556. — For a description and a critical discussion on the approaches of Michels see: Pfetsch, Frank, *Robert Michels als Elitentheoretiker*, in: *Politische*

Table 4: Configuration-Sequence-Analysis (Computed by SPSS)

Code	columns	format	variable
o1 - o2	F2.0		Beginning First Position (Year)
o3 - o4	F2.0		Finish First Position (Year)
o5	F1.0		First Position
o6 - o7	F2.0		Beginning Second Position (Year)
o8 - o9	F2.0		Finish Second Position (Year)
1o	F1.0		Second Position
			... Six Positions

*SPSS-Procedure*

Compute Sequenz = (Posit 1\*10\*\*5) + (Posit 2\*10\*\*4) + (Posit 3\*10\*\*3)  
+ (Posit 4\*10\*\*2) + (Posit 5\*10) + Posit 6

Frequencies General = Sequenz ...

*Configuration-Tableau: The Case of Trade Unions Employees*

Configuration-sequence	Total Populat.	Election 1898	Election 1903	Election 1907	Election 1912
1	86=67.2 %	8=61.5 %	27=61.5 %	35=68.6 %	59=64.1 %
10709	1	1	0	0	0
12	6	0	2	2	5
123	1	0	0	0	1
13	7	3	2	3	2
1313	1	0	0	0	1
1351	1	0	1	1	1
138	1	0	0	0	1
14	14	0	2	3	13
15	3	0	1	1	3
156	1	0	0	1	0
16	1	0	0	0	1
17534	1	0	0	1	1
1794	1	0	0	1	1
18	1	1	1	1	1
19	1	0	0	1	1
195	1	0	1	1	1
Total	128=100 %	13=100 %	36=100 %	51=100 %	92=100 %

Vierteljahresschrift 7 (1966), pp. 208 ff.; Ebbighausen, Rolf, Die Krise der Parteiendemokratie und die Parteiensoziologie, Eine Studie über Moisei Ostrogorski, Robert Michels und die neue Entwicklung der Parteienforschung, Berlin 1969; Röhrich, Wilfried, Robert Michels. Vom sozialistisch-syndikalistischen zum faschistischen Credo, Berlin 1972.



In closing, allow me to express the hope that through a methodological reorientation of the analysis of collective life-histories the one-sided bond of question formulation will be broken from the empirical research of elites and a more comprehensive formulation will be established.

### Abbreviations

Abbrev.	Code	Meaning
KA	0	„Kein Arbeiterbeamter“ (the candidate had not been employee of working-class movement up to 1914)
GA	1	„Gewerkschaftsangestellter (Lokal und Regionalebene)“ (trade union employee with local and regional functions)
LGA	2	„Gewerkschaftsangestellter (Leitungsfunktion)“ (trade union employee with leading functions)
AS	3	„Arbeitersekretär“ (generally a trade union employee with most important local functions: representative of all local unions and chief of an office for advertising the workers)
PS	4	„Parteisekretär“ (party secretary)
PR	5	„Parteiredakteur, Parteischriftsteller, Berichterstatte“ (party editors, writers, authors, reporters)
LPA	6	„Parteiangestellter (Leitungsfunktion)“ (party employee with leading functions)
PA	7	„Parteiangestellter (ohne Leitungsfunktion)“ (party employee without leading functions)
GNA	8	„Genossenschaftsangestellter“ (Co-operative employee)
KA	9	„Krankenkassenangestellter“ (health insurance employee)

*Example:*                    Meaning of sequence-configuration „123“  
 „123“ describes the career of a working-class movement employee: first he was occupied as GA (First Position), then he was occupied as LGA and finally he was occupied as AS (Last Position).



Fig. 2: Frequency Distribution of the Sequence „First Position/Last Position“

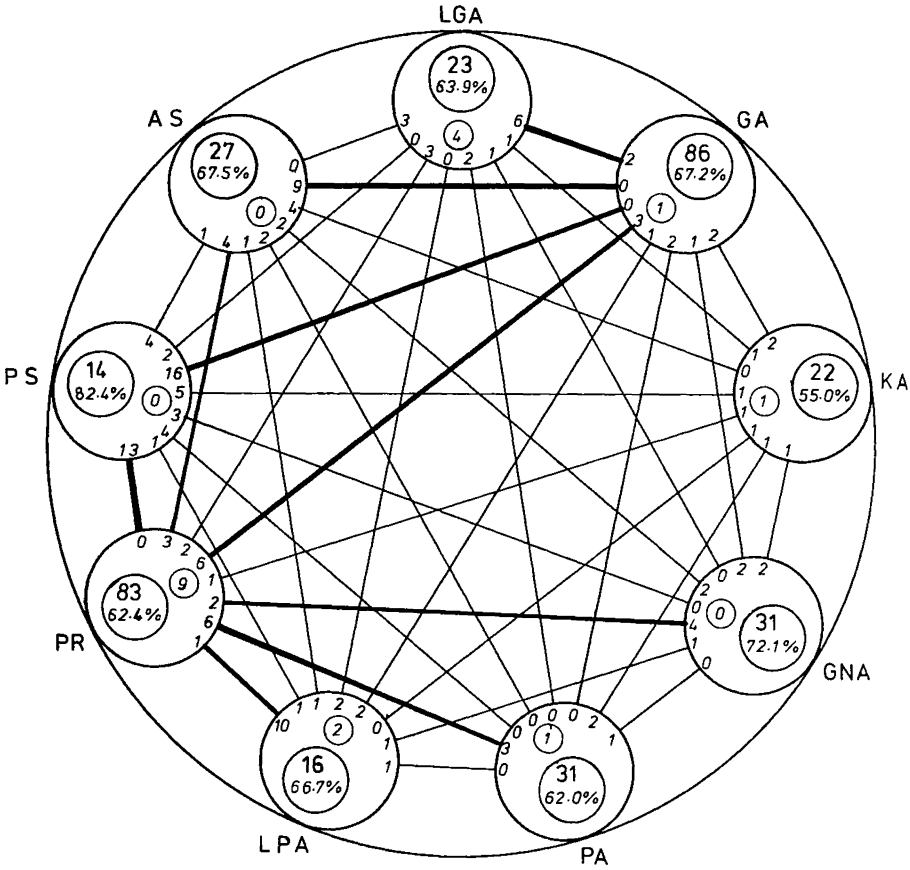
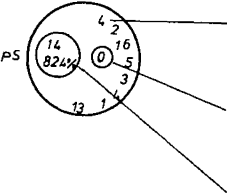


Fig. 2:



Frequency of configuration-sequences if the last position is equal with the special position:  
4 „Arbeitersekretäre“ (= First Position) finished their career as „Parteisekretäre“

Frequency of self-recruitment:  
There ist no „Parteisekretär“, who began as a „Parteisekretär“ and finished the career (the configuration should have at least 3 positions) again as „Parteisekretär“

Frequency of immobile „Parteisekretäre“ (only one position of career):  
14 candidates began as „Parteisekretäre“ and remained „Parteisekretäre“

Fig. 3: Configuration-Sequence of Trade-Unions-Employees („First Position“ = „GA“)

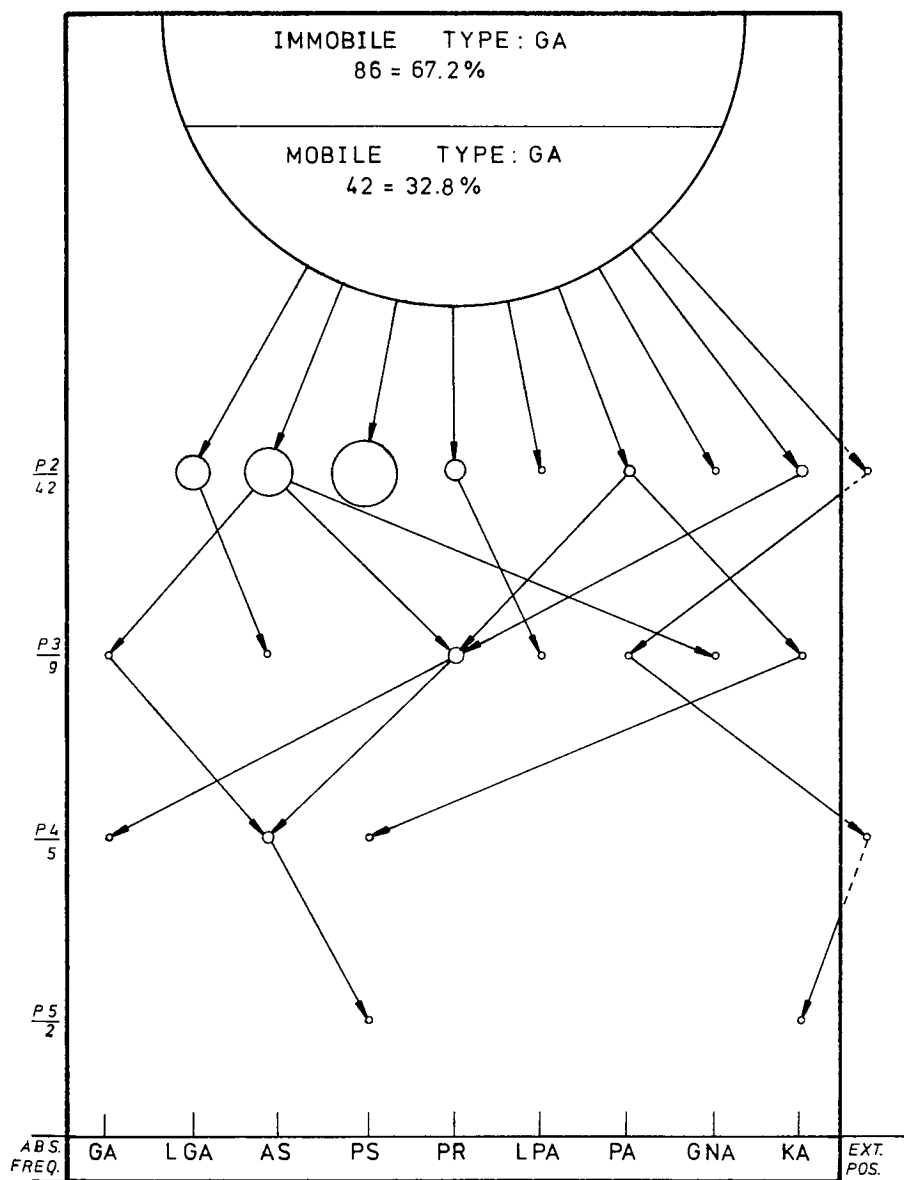


Fig. 3:



diameter of circle = number of cases

Position 1: 14 „Gewerkschaftsangestellte“ (GA) became „Partei-sekretäre“ (= 2. Position)

Fig. 4: All Candidates, Candidates 1898 and Candidates 1912  
Compared by the First Year of Career-Entrance

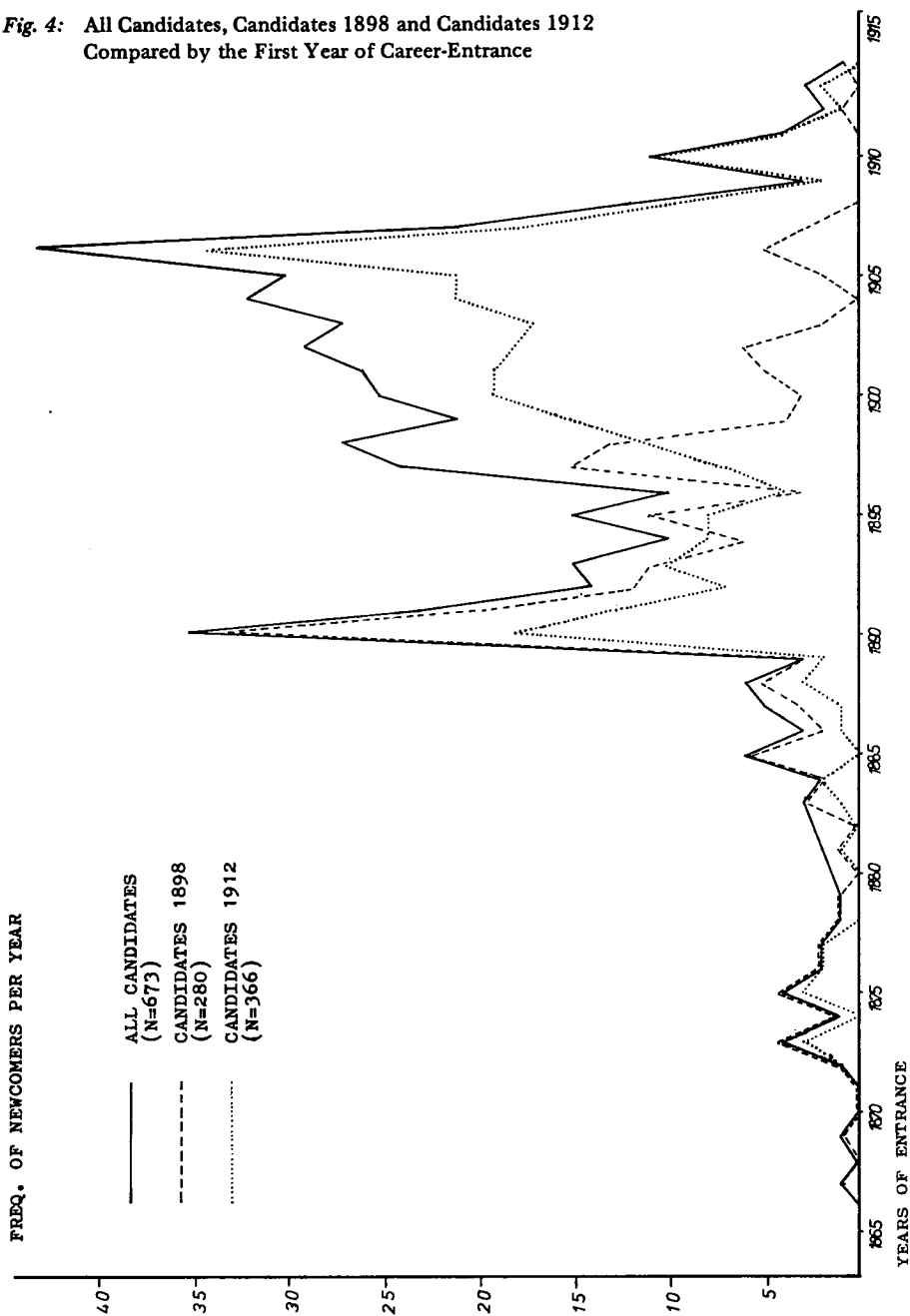
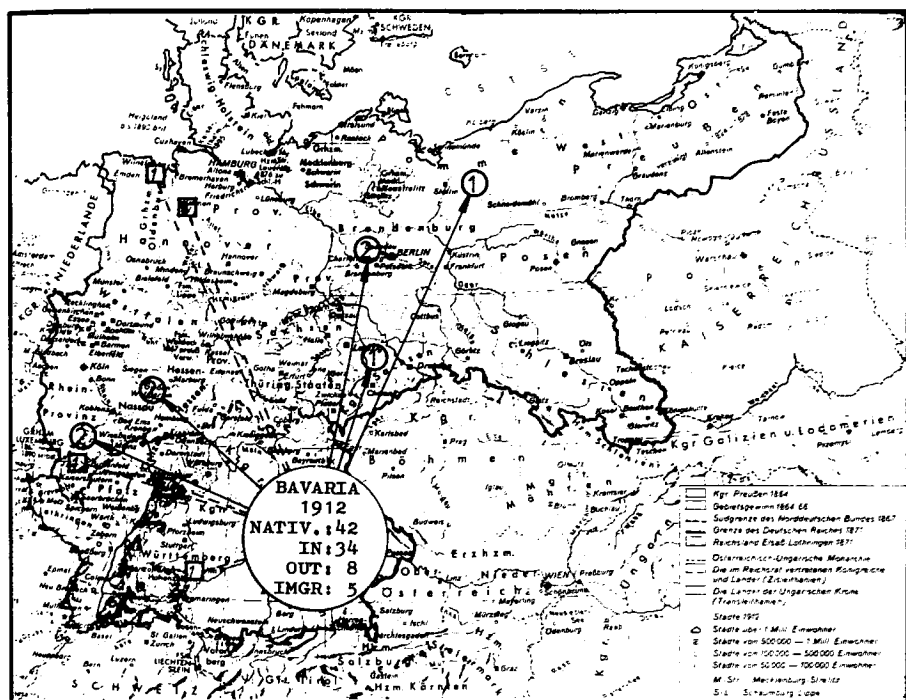


Fig. 5: Regional Mobility (Place of Birth vs. Place of Residence)  
The Case of the Bavarian Candidates 1912

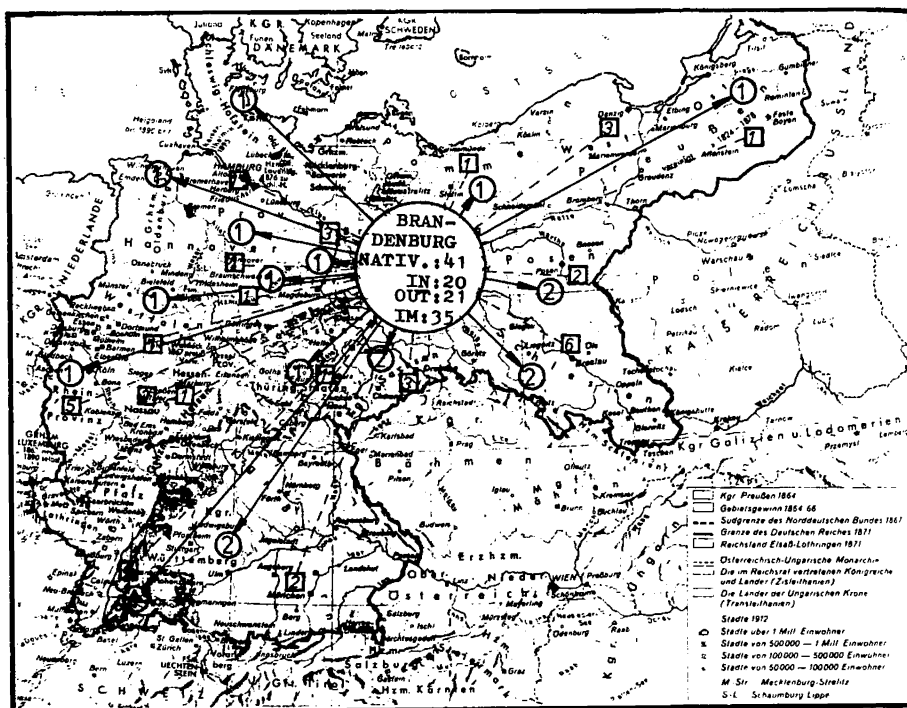


Map from: F. W. Putzger, Historischer Weltatlas, 98th ed., Berlin 1974, p. 97 (Mitteleuropa 1866–1914)

Explan.: 42 candidates were borne in Bavaria. 34 Bavarian candidates lived still in the region of their birth. Only 8 Bavarians were mobile and worked in other regions of Germany. Only 5 Non-Bavarian candidates lived in Bavaria.

- Freq. of Bavarian „emigrants“ per region
- Freq. of Bavarian „immigrants“ per region

**Fig. 6: Regional Mobility (Place of Birth vs. Place of Residence)  
The Case of the Brandenburg Candidates 1912**



**Explan.:** 41 candidates were born in Brandenburg (centre: Berlin).  
20 candidates lived still in the region of their birth.  
21 Brandenburg candidates were mobile and worked in other German regions.  
But 35 Non-Brandenburg candidates lived in Brandenburg.

- Freq. of Brandenburg „emigrants“ per region
- Freq. of Brandenburg „immigrants“ per region

## The Use of Collective Biography in Research on the Imperial Russian Civil Service

In the light of the central importance that the state and its officials have had from the very earliest period of Muscovite history, it may appear surprising that only during the past ten to fifteen years have 20th century historians devoted serious attention to problems relating to the nature and development of officialdom. The explanation for this neglect is, however, relatively simple. Since the revolution of 1917 scholars, both within and outside of the Soviet Union, have been preoccupied with that event. Emphasis, from one perspective or another, was placed upon the worker and peasant movements or the condition of those groups, on the development of radical ideologies, or on the problems and ultimate failure of moderate views and programs. Soviet scholars, except those dealing with very early periods, were not interested in the discredited institutions of Tsarist autocracy, and Western scholars were more concerned with why liberalism did not succeed, than in finding out how the Tsarist regime and society actually functioned for half a millenium.

This situation has changed markedly in the past ten to fifteen years with the appearance of an important number of studies, both in the U.S.S.R. and in the West, which have been ably discussed by Professor Orlovsky in an excellent review article<sup>1</sup>. These studies have dealt with both the structure and operation of Tsarist governmental institutions and with the personnel who staffed them. It is in the studies dealing with staffing that quantitative methodology, namely collective biography, has come to be employed. Before passing to the quantitative study of official careers we should note, however, that the institutional, legal and ideological milieu of officialdom has been exhaustively analyzed by Torke for the first half of the 19th century but for other periods we lack studies of comparable quality and depth<sup>2</sup>.

On the quantitative side, serious work has been done on the 17th, 18th, 19th and early 20th centuries, for the bureaucracy as a whole, and to a lesser extent for important parts of it, such as the main central agencies, provincial agencies, and also

<sup>1</sup> Orlovsky, Daniel T., Recent Studies on the Russian Bureaucracy, in: *The Russian Review*, 35 (1976).

<sup>2</sup> Torke, Hans J., *Das Russische Beamtentum in der ersten Hälfte des 19. Jahrhunderts*, in: *Forschungen zur osteuropäischen Geschichte*, 13 (1967).



for specific ministries such as Justice, Interior, or State Domains. Some work has encompassed all levels, from the highest to the lowest; other studies have concentrated on elite groups of one kind or another.

Why has there been such an extensive development in this field in recent years? In part it is due to the general increase in the use of quantitative approaches to historical studies, but in the case of Russian history it primarily reflects the existence of a gaping hole in our knowledge of Russia, and, like nature, historians abhor a vacuum. Little more than a decade ago professional Russian historians had no solid bases for statements they made about the official class. It is hardly an exaggeration to say that many had no better knowledge of the subject than a casual reader of Gogol's famous story, *The Overcoat*, whose pathetic hero, Akakii Akakievich, had come to personify the Russian official in most people's minds for lack of any alternative image. The second major reason for the rapid progress of quantitative studies of Russian officialdom is that, once historians began to look for it, very large amounts of data were to be found. Russia has had a centralized, record-keeping government at least since the fifteenth century and the close relationship of state service to social status and what eventually came to be recognized as the nobility ensured that vast amounts of paper and ink were used in establishing the credentials of state servants.

Of course, by no means all of the documents created have been preserved, but enough survive to permit serious work to have been undertaken at least from the early 17th century forward and possibly even earlier. Officialdom was not the creation of Peter the Great, but in the 18th century the number of officials greatly increased, and so too does the availability of records dealing with them. In the 1750's the Empress Elizabeth ordered a full census of all officials and the results of that survey have recently been analyzed by the late S. M. Troitskii in a study published in Moscow in 1974, the first major Soviet work on the Imperial period to make extensive use of the life histories of officials<sup>3</sup>. The census of officialdom of Empress Elizabeth was a one-time thing, never to be repeated, but in the late 18th century more methodical and extensive records began to be kept by the Heraldry Office, the agency charged with determining who was, and was not, a member of the nobility. From the 1790's on all agencies were obliged to supply the Heraldry Office annually with a full description of each of their employees, a description that included substantial information, not only on the official career but also the general life history of the official. These were the *formuliarnie spiski* or personnel records. Most of them have perished over the years, but enough have survived to permit extensive research on many aspects of officialdom in the first half of the 19th century. The Heraldry Office was abolished at mid-century, but the *spiski* continued to be compiled in the same form as they had been in the 18th century. In the second half of the 19th century, most of the information contained in them was published in lists of officials, sometimes arranged by rank and sometimes by agency. In short,

<sup>3</sup> Troitskii, S. M., *Russki absolutium i dvorianstvo xvii v, formirovanie biurokratii*, Moscow 1974.

extensive data are available on Russian office holders down to the revolution of 1917, although there is no guarantee that a particular agency in a particular year will be available, particularly in the first half of the century.

What do these personnel files contain? In good bureaucratic fashion more and more information was required as the years went by. From the 1840s on the amount of data included is particularly impressive. Personal information includes name, age, religion, marital status, name of wife if living, number and ages of children, religion but not social origin of wife, detailed information on real property, serfs, uninhabited rural land, and urban real estate. They even specify whether houses are stone, wooden, or wooden with stone foundations. Furthermore, property holding is broken down by husband and wife, and in terms of method of acquisition, inherited or acquired and includes the holdings of the official's parents, if they were living. The only item of personal data that is not included is place of birth, even the location of the estates listed is rarely given. This omission surely is a result of the long-standing centralization of the Russian state and the absence of local loyalties among the nobility.

However, the main function of these personnel files was not to present economic and demographic data, but to record the service of the official to the state. Here the record is extremely detailed. It usually begins with the highest level of education reached by the individual and then proceeds to record every position, promotion, award, and pay raise he received in his entire career. If he served in the army at any time special attention is given to any occasion when he „came into direct contact with the enemy“. There is also space to record all extended leaves of absence and finally the official's current rank, job, and pay, including allowances for housing and meals, if any are included. The record of a senior official with many years of service can easily amount to a booklet of 40 to 50 large folio pages all carefully written out in elegant script. Even a junior clerk's file fills four or five pages. Since every agency was required to submit a new set every year, for every employee included in the table or ranks, one can immediately see what many of the clerks did all year long; they copied out last year's service records, adding a year to the official's age and including any new posts he had achieved.

If one is interested only in the officials serving in the nineteenth century, one could hardly ask for more detailed information, and it is available for a wide variety of agencies, both central and provincial, although not for all. The only major lack in terms of the questions likely to interest Russian historians is any indication of how long standing the official's membership in the nobility was. A man listed as „from the nobility“ could be the son of commoner ennobled through service or from a noble family of many generations. There are ways to approach this problem, but none are entirely satisfactory.

Historians, however, by their very nature are inclined to look at change over time, and here very significant problems arise when using the *spiski*. Those compiled prior to the 1840s, although the same in basic format and purpose, are far less detailed. For the social historian the most serious omissions are the property holdings of the official's parent, and the education of the official prior to his entrance

into service. When parent's property holding is omitted it becomes impossible to separate the son of wealthy parents who has yet to inherit the family estates, from the landless noble, of which there were many. Troitskii, working with the 1755 census of officials had even greater difficulty, because 22 percent of his officials did not even report their social origin. Through elaborate and exhaustive examination of each individual case (some 1,214 in all) he was able to determine their probable social origin with reasonable certainty, a task possible only for a scholar with continuous and unrestricted access to the archival sources<sup>4</sup>. Western scholars who have the opportunity to use archival material for necessarily restricted periods of time must turn to other strategies or do without.

What kinds of questions have been dealt with on the basis of this material? We can examine the nature of the official career at different times. Was it purely civil? Mixed civil and military? What kind of movement among agencies was there in the course of a career? What training, if any, preceded actual employment and at what age did men typically start their careers? Without going into details we can say that it has become clear from recent work that in the course of the century from 1750 to 1850 the character of the official career in Russia fundamentally transformed, a fact that could only have been established through extensive use of quantitative materials. Legal sources, readily available in the *Complete Code of Laws of the Russian Empire*, at times represent an expression of what policy makers desired, and in other instances, a ratification of developments that were already in effect. An excellent example of the latter phenomenon has developed through the examination of career records which show clearly that formal, institutionalized education became the norm for civil officials well before it was required by law in the early 19th century.

A second category of problems relates to the personal background of the official, his social origin, religion, and economic position. There are a host of obvious questions that spring to mind. Were high-ranking officials always nobles; if so, were they landed nobles? Did non-nobles who rose in service to high-ranking posts acquire landed wealth? Were certain agencies the preserve of nobles and others not? Did social origin affect career success for men with comparable education? And so forth. All of these matters and many similar ones have been dealt with on the basis of data derived from the service records in an attempt to locate the official in the totality of the Russian social system<sup>5</sup>.

The third category of data, that relating to material status and the numbers, sex, and ages of children has, as yet, been little used except in connection with the property holdings of officials' wives. The data on children, which are extensive in mid-nineteenth century files, have not even been collected. For someone with experience

<sup>4</sup> Troitskii's discussion of the problem is some seventy-three pages long. Troitskii, op. cit., pp. 180-253.

<sup>5</sup> These questions and many others will be dealt with in the forthcoming volume: Pinter, Walter M., and Rowney, Don K. (eds.), *Russian Officialdom: The Bureaucratization of Russian Society from the Seventeenth to the Twentieth Century*, Chapel Hill 1979.

in family history or demography this material may present opportunities that are not apparent to those primarily interested in social and administrative history who have worked with this material thus far.

The methods that have been used in collecting and analyzing the data from the service records have been relatively straight-forward and little attempt has been made to apply highly complex or sophisticated statistical techniques. The actual collection of the data, if archival material is involved, is simple but extremely time-consuming. The records are arranged so that the basic social, economic and demographic data can be quickly noted from the first two pages. However, recording career data such as the date of entrance into service, the date of achievement of specific ranks, receipt of awards, transfers from one agency to another, and so forth can frequently require reading through many pages of information. Details on careers have therefore been collected only for relatively small groups of high ranking officials. For large studies involving thousands of officials the data used have been limited to current rank and agency of employment and the date, level, and agency of first employment.

The processing of material has been carried out in the manner now familiar to most historians concerned with quantitative materials. The information is coded, put on punch cards and then, usually, on to tape for analysis via a computer. The basic technique used in working with the data has been cross-tabulation, broken down in innumerable ways by using different sub-groups. Even the calculation of such a simple statistic as the average age of officials is aided immeasurably by the computer when thousands of cases are involved and many sub-groups are being compared. To a limited extent, scholars using the data on Russian officials have employed other statistical techniques, correlations and analysis of variance and covariance. The limited use made of more sophisticated statistical techniques can perhaps be explained in many instances by the lack of statistical sophistication of the researchers. But it is certainly not true of everyone in the field. It seems to be true in most instances so far that the important conclusions can be derived from the data through careful working over of the material using only the most elementary tools. The more advanced approaches have, as yet, only conformed and stated in more concise form conclusions already evident from more simple-minded approaches. This does not mean that there is no room for high-powered statistical tools in the study of the life-histories of Russian officials; there may well be, but they clearly have not played a major role thus far. One great advantage of the present state of affairs is that everything written about Russian officials remains accessible to all historians, not only those with extensive training in statistical analysis.

Some of the specific problems that have come up thus far may illustrate both opportunities and the limitations inherent in the kind of data that are in the service records. Perhaps the most serious problem for an historian attempting to examine trends over any extensive period of time is, as already mentioned, the lack of important data in the records for the earlier periods. To a significant extent this problem can be overcome by grouping officials according to the date they entered service, when using the extensive information available on men in service in the

mid-nineteenth century. Thus, several thousand officials serving in the 1850's can be arranged in „year of entrance cadres“, and these cadres can then be compared in terms of important variables such as inherited property, education, and so forth, which are not available in the service records compiled early in the nineteenth century. Since men began service in their teens at the beginning of the century and served till their sixties one can go back about forty years using this technique. It does, of course, bias the sample in favor of the long-lived and those who do not retire early.

Another way to deal with the problem of missing information is to find some other kind of data that can effectively act as a substitute. Neither nationality nor place of birth were recorded in the service records. It is usually possible, however, to recognize the foreign born from other features of their service record, their place of education, or the like. However, the major group of non-Russians in Imperial service were from the minority groups within the empire, particularly the Baltic Germans and the Poles. One possible way of identification would be through their names, but this approach presents serious difficulties because of the very substantial amount of cultural „Russianization“ that had taken place over the centuries during the expansion of the empire. Nevertheless, if a better alternative had not been available, it could have been attempted. Fortunately, religion was consistently reported on all service records and it is a reasonable assumption that the Lutherans were mostly Germans from the Baltic provinces and that the Catholics were Poles. Some of the orthodox may well have been of German or Polish background, changing their religious designation for reasons of convenience or of conviction. Therefore the use of religion as a substitute for nationality may tend to undercount the non-Russian element to some extent. It would be possible to classify officials in terms of their names' presumed national origin and see how well that grouping matched one made on the basis of religion. This would produce an estimate of the orthodox officials of non-Russian background. The effort involved for a large scale attempt of that kind does not seem justified. In the case of a detailed examination of a single ministry it might well be worthwhile.

Except in the relatively rare instances where data on an entire population are both available and of manageable size, the researcher is faced with the problem of the „typicality“ of the material at his disposal. The techniques of random sampling are well established and can be applied in cases where total populations are available. Historians, however, are frequently confronted with „accidental samples“ and the collection of service records in the Central State Historical Archive in Leningrad is a perfect case in point. Some time in the late 19th century most of the enormous mass of service records (every employee, every agency, every year) that had accumulated for perhaps seventy-five years was discarded. Clearly nobody ever consulted them. When you turn the pages on the surviving examples, the sand used to blot the ink falls off and piles up on the table. There is no apparent rhyme or reason to what survives among 22,716 individual items that remain, according to the official archival guide. Many of them are files on a single individual. Others are huge volumes nearly two feet thick with all the files of a ministry for a given year, but

sometimes only volume one of a two or three volume set survives. Someone started to put the individual service records into alphabetical order by the officials' surname, but without regard to date or agency of service. Fortunately for contemporary quantitative historians this misguided effort was abandoned only a few letters down the alphabet from „A“.

There do survive a considerable number of files for whole ministries or at least departments of ministries. For a given agency one can then deal with a total population. Sampling is not indicated, for the numbers are not that large. But if one is interested in making generalizations about Russian officialdom as a whole there is no statistical way one can assess the typicality of a given group of agencies. One must rely on qualitative data and historical common sense. If, for example, you are interested in the prevalence of retired military men in the civil service and you have data for several civil departments, but not the Ministry of War, the figure you get must be regarded as a probable minimum. If you have a fairly diverse group of agencies and, with respect to some variable, they are very similar, the likelihood is strong that the agencies you do not have will also be similar failing any particular reason to the contrary.

This is all very obvious, but the point is that all the data in the world about the Ministry of the Interior will tell you nothing about the civil service as a whole unless you have some knowledge of that ministry and the other agencies of the day that permits you to evaluate the data. A similar problem arises when making comparisons over long periods of time. Agencies come and go, and, to some extent, so do actual functions. „Comparable“ groups of agencies fifty years apart can, at best, be approximate.

A special problem in the case of the Russian civil service, but one that may have its counterpart elsewhere, is the existence of parallel hierarchies of ranks and offices. Ranks, conveniently numbered from 1 to 14, are enticingly easy to use quantitatively because every official had one and it is clearly listed on his record, along with all his prior ranks, dates he achieved them, and so forth. The hierarchy of offices is less easily accessible. Jobs are named or described in the individual's service record, but to determine the job's place in the hierarchy one must turn to the legislation establishing the agency involved. Nevertheless, particularly for the late 19th century, it is becoming clear that the hierarchy of offices was functionally more important than that of ranks and scholars are beginning to deal with this question.

Finally, and for much of the research on the Russian civil service, the most serious problem is how can the social origin of an official be determined more than one generation back? The service records indicate the status of the official's father but nothing more. Given the nature of the Russian system it is of crucial importance to make some judgment as to how many sons of nobles are sons of long-term nobles, and how many of recently ennobled men. With small elite groups it is possible to undertake careful geneological and biographical research, going beyond the simply service record. Professor Meehan-Waters has done this exhaustively for the *Generalitet* of the 18th century<sup>6</sup>. For anything but the most restricted groups this is a terribly tedious process with dubious prospects of success. The basic biographical

encyclopedia for Imperial Russia was compiled in the late 19th century, and some volumes of the alphabet were not complete before the revolution and were never published. Distinction in government service was only one of many reasons for inclusion so it tends to bias a group drawn from it toward those with distinction in science or the arts, men for whom government service was possibly a secondary or incidental career. Meticulous archival research like that of Troitskii over a long period can provide much of the missing information but that is extremely difficult for non-Soviet scholars to arrange. The best one can do in terms of overall judgment on the mass of officialdom is to look at the other data available and make reasonable deductions from it. In this case the information on property holdings shows the existence of a large group of officials from totally landless families, strongly suggesting that many, if not most, of them were descended from relatively recently ennobled civil officials or military officers.

What are the major tasks still undone using collective biography techniques as a way of studying Russian officialdom? As mentioned above, the data on marriage and children have not been seriously examined by qualified people. In the area of administrative and social history we now have a fairly good general picture of the nature of Russian officialdom from as early as the 17th century down to the revolution of 1917 and, at least to some extent, into the Soviet period. What is only beginning to be done, notably by Professors Rowney and Orlovsky for the Ministry of the Interior, is to look carefully at the staff of specific agencies over a relatively short period of time and begin to relate changes in personnel, patterns of promotion, transfers in and out of service and so forth to government policy and other major events in the society at large. We are not interested in the nature of the civil service for its own sake but because it was part of a larger picture of Russian development. The collective study on the life histories so diligently and massively recorded collected by generations of Russian clerks is only a beginning, but an essential one, for solidly based understanding of what the Imperial Russian government actually was, and what it could and could not do.

<sup>6</sup> Meehan-Waters, Brenda, *The Muscovite Noble Origins of the Russians in the Generalitet of 1730*, in: *Cahiers du monde russe et soviétique*, 12 (1971).

## IV. Content and Document Analysis





## Computer Aided Content Analysis of Historical and Process- Produced Data: Methodological and Technical Aspects

Document analysis as an intellectual as well as a computer aided procedure has a sufficiently long tradition to be recognized as a standard instrument in social research. Like interview and observation, content analysis is used to generate data for the analysis of social reality. Records from almost any source of communication can provide an empirical base for statistical analysis and subsequent interpretation. Content analysis, in particular the computer aided procedures, will be reviewed here as a possible candidate for inclusion in the set of instruments that will be needed. For this purpose an attempt is made to answer three questions:

- 1) What can quantitative history expect from content analysis?
- 2) What computer aided procedures are available?
- 3) How can content analysis be applied to historical and process produced data?

Suggestions for introducing this method into the research process dealing with historical and process produced data will be made. The final decision on whether and how it can be used will depend, of course, on the objectives of substantive research.

### 1. What Can Quantitative History Expect From Content Analysis?

Content analysis may well be the appropriate instrument for Social Science History whenever „textual data“ not containing quantitative information per se (like statistical tables, turnover figures etc.) are under investigation. Text books on social science methodology recommend content analysis as the obvious method for research on communication. These can be communications from manifold sources

<sup>1</sup> Definitions of content analysis are discussed in e. g.: Markoff, J., et al., *Toward the Integration of Content Analysis and General Methodology*, in: Heise, D. R. (ed.), *Sociological Methodology* 1975, San Francisco 1974, pp. 4–7, and Holsti, O. R., *Content Analysis for the Social Sciences and Humanities*, Reading 1969, p. 2 ff.

like speeches, pictures, movies and other manifestations of symbolic behavior. In this contribution we will confine ourselves to written documents.

Without discussing the many definitions of content analysis we propose to label all procedures content analysis that comprise the systematic description, reduction and inspection of communication under the analytic frame of research concepts<sup>1</sup>. Subsequent analysis of the relations between these concepts may include inferences on origin, context and reception of the communication.

Part one of this procedural definition covers the aspect which attributes the distinctive feature to content analysis as compared to other instruments. As the process of interviewing generates data from responses, content analysis does the same when applied to texts. Frequently content analysis projects stop after displaying the frequency distributions for these data. It is obvious that it would be inappropriate to consider research finished after textual data is converted to numerical form. So far it has achieved not much more than information reduction. Analysis and interpretation have to start thereafter. Looking at content analysis in this way may present a subtle misunderstanding of the value of frequency counts on concept or word occurrences. They are as much a final result of research with content analysis, as the marginal distributions are the final result of research employing interviews. It is a question of research design of how much you get out of it by further analysis.

Before addressing this area, however, we shall recall in what ways reduction of text can be achieved.

It has been postulated that coding text by content analysis procedures should be semantically as close as possible to the contents of the original documents<sup>2</sup>. This basically means taking redundancy out of the text and boiling it down to its meaning constituents. If, in addition a standard descriptor language is used, heterogeneous information from the original representation becomes comparable. In this way, similar cases can be grouped together and thus can be counted. The intention of this approach is not to read between the lines or to involve analytical processes in the information reduction step. It is rather a condensation of verbose description into statistics as a result of a coding and counting process. Analytical and inferential steps may be based on these results, no longer on the original information. The question must then be asked whether after this kind of reduction instrumental content analysis is still possible. For representational approaches it may be the appropriate procedure. But content analysis can go beyond condensed semantic representation. It can be analytic in the reduction process. This is usually the case in traditional content analysis. Psychological applications for instance, have shown that it is possible to capture the pragmatic aspects when the coding process builds on the connotative meaning of words or the latent meaning of larger communication units.

Traditional content analysis, employing human intellect for the reduction process, requires the translation of the text under investigation into a predefined cate-

<sup>2</sup> Markoff, *Integration*, p. 3.

gory scheme. The categories representing the research variables are defined intentionally. A coding rule describes which set of denotation units from the text should be grouped under one category. The more familiar the human coder is with both the text under investigation and the classification scheme, the better the expected result. This is optimally the case, if a native speaker, familiar with the subject area, is conversant with the classification scheme as well.

But what performance can be expected from a coder who is not familiar with the context of a given text? He will hardly manage to understand the text entirely unless it is explicit and simple. If the text contains insider information, he will be lost on these grounds, even though he may have above average linguistic competence. He may not be able to get the latent or even the manifest meaning. Nevertheless he may well succeed in breaking down the text into smaller communication units and assign categories to them. This may be a problem of particular importance in the attempt to analyze historical documents when context information is lacking, whereas process produced data as a rule have, by administrative law, a clearly defined context.

Computer aided procedures are not able to handle complete texts as entities nor are they able to correctly identify the boundaries of meaning units aside from syntactical boundaries. We consciously exclude experiments in artificial intelligence which achieved some practical results, however, in a small, well defined domain. Thus in most applications the text is broken down into the single words by the input program. The subsequent programs then relate the denotative and connotative meaning of the individual words from the text to the categories defined a priori in a dictionary or they ascertain statistical associations and measures in an empirically exploratory way.

## 2. Which Computer Aided Procedures Are Available?

As suggested above, we may distinguish the following procedural steps of a content analysis: 1. Description, 2. Reduction, 3. Inspection. Along these lines, we will describe what options are available in the various software systems and what they can be used for.

### *2.1 Descriptive Procedures*

Automatic procedures process text as characters or character strings. Units which are separated by empty spaces or other delimiters are recognized as words. Thus most systems operate on the single word as the unit of information. The listing of

all words occurring in the text in different forms (token) as well as frequencies of identical word forms used (types) and alphabetically sorted indexes can be generated easily. Most programs generally available allow modifications of these lists according to ascending or descending frequency and forward or backward alphabetization. These frequencies allow the computation of various quotients like Type Token Ratio (TTR). The TTR is used for the characterization of the differentiated word usage or the restricted word spectrum of the text source. Drawing on extralinguistic information, e. g. stress at time of the generation of the message, the TTR can be used as a measure of stress intensity by relating the actual figures against the individual standard and computing the deviation. Various TTR computations for the same text source have proved to be relatively constant. Like Markov transition probabilities in the sequencing of words or arguments, they can be used for authorship identification. To support disambiguation and dictionary construction, Key Word in Context routines (KWIC) can be used. They list the words after permutation in their textual environment. KWIC, as well as KWOC (Key Word out of context) are available at almost every installation that offers text processing facilities<sup>3</sup>.

## 2.2. *Information Reduction by Categorization*

Empirical social research applies particular measurement techniques and categorizations to describe social reality. Interview surveys are used to collect data describing properties of individuals. The goal of measurement is the grouping of comparable characteristics in answer categories of variables, which then can be statistically analyzed to identify interdependencies between the variables. Applied to content analysis procedures, this categorization process can be characterized as the grouping of word occurrences (token, individual characteristics of the particular text) under stem forms (types, characteristics as combination of different forms), which can then be assigned to theoretical categories on a higher level (variables). Alternative procedures used for categorization will be exemplified by drawing on the routines in the most prominent system, the GENERAL INQUIRER<sup>4</sup>.

<sup>3</sup> At most universities the department for Linguistics will have suited software. For special developments see also: Genet, J. G., *Medieval History and the Computer in France*, in: *QUANTUM Information*, 5 (1978), pp. 3–10.

<sup>4</sup> Stone, P. J., et al., *The General Inquirer: A Computer Approach to Content Analysis*, Cambridge/Mass. 1966.

### 2.2.1. A Priory Categorization by Dictionaries

The GENERAL INQUIRER system assigns text entries to theoretical categories by various program steps. The relation between the possibly relevant entry words and the categories is predefined in a content analysis dictionary. About 30 different dictionaries are in existence<sup>5</sup>. The more recent dictionaries contain context information for disambiguation of homographs<sup>6</sup>.

Not all words occurring in the text base can be anticipated when creating the dictionary. These 'new words' will be displayed in a leftover list. On this basis the investigator can make few tag-entry assignments and incorporate them into the dictionary. Results of the tagging operation are displayed in the TAG TALLY LIST. It gives frequencies in absolute figures and percentages on the basis of the total number of words and sentences in the text. The results are stored on magnetic tape or disk for post-tagging operations, such as retrieval of particularly interesting text parts for closer inspection, further analysis of purposely selected subsamples, and interfacing to statistical analysis packages.

The principles of the dictionary approach developed for the GENERAL INQUIRER were adopted by systems like EVA, TEXT, TEXTPACK and SPENCE's<sup>7</sup>. While the idea of the GENERAL INQUIRER was to offer a general content analysis instrument, the more recent developments were initiated for special application needs. EVA was developed for the analysis of newspaper headlines, ANACONDA and TEXTPACK were developed for the coding of answers to open ended questions, and SPENCE's programs for the analysis of psychiatric interview protocols. Since they are special purpose oriented, they developed certain special features further while neglecting others that are needed for more general applications.

TEXTPACK offers a very efficient set of routines for dictionary comparisons, text correcting and selection of particular answer texts. EVA was aiming at further developments for an advanced semantic analysis of headlines. Both systems offer direct interfaces to statistical analysis packages (EVA — RAPROSYS, TEXTPACK — OSIRIS). The SPENCE programs are particularly suited for smaller texts (up to 100 lines per segment). Even though COCOA, written for linguistic analysis, does not offer tagging routines it is appealing to use for teaching basic procedures in content analysis<sup>8</sup>. It offers a very flexible parameter language for structuring and labeling the input as well as all sorting and index list options.

<sup>5</sup> An overview of General Inquirer Dictionaries (status: Fall 1965) is given in Stone, op. cit., pp. 140–141.

<sup>6</sup> A detailed description is given in: Kelly, E., and Stone, P.J., *Computer Recognition of English Word Senses*, Amsterdam 1975.

<sup>7</sup> Methods and techniques of available content analysis software are reviewed in: Mochmann, E., *Automatisierte Inhaltsanalyse*, in: Langenheder, W. (ed.), *SIZSOZ Expertisen: Ausgewählte Gebiete sozialwissenschaftlicher DV-Anwendung*, Vol. 1, St. Augustin 1976, pp. 61–92.

<sup>8</sup> Berry-Rogge, G. L. M., and Crawford, T. D., *COCOA Manual*, Chilton, Didcot, Berkshire. A tagging routine has been added in Cologne.

### 2.2.2. Inductive Categorization by Statistical Procedures

Parallel to dictionary systems, the empirical inductive procedures were developed. They explicitly avoid an a priori categorization<sup>9</sup>. The criticism of dictionary approaches can be subsumed under the following arguments: A priori dictionaries are derived from or oriented to sociological or psychological theories. Thus they reflect particular research intentions. Empirical approaches are neutral in that respect. In addition there are serious doubts whether dictionaries can appropriately anticipate the vocabulary domain of the sociolinguistic community that produced the text under investigation.

The leading, and so far only inductive system, is Iker's WORDS system<sup>10</sup>. The conceptual design of WORDS was based on the intention to exclude any influence of the researcher on the derivation of concepts from the text. The themes and theoretical concepts should be derived from the texts by means of statistical procedures. The unit of information is the single word, based on the assumption that sufficient meaning resides in the words and in the associations among and between words to produce an accurate representation of content. The procedure can roughly be characterized by the following steps:

- 1) An input document is divided into „segments“, e. g. the page, the paragraph, equal numbers of words, time units in order to achieve comparable length of documents;
- 2) all function words, e. g. articles, conjunctions, are removed;
- 3) remaining words are deinflected to root form (lemmatized);
- 4) the frequency of occurrence in the segment is computed for each different word;
- 5) a subset of these words (types) is selected for analysis;
- 6) an intercorrelation matrix (ICM) is computed on this selected subset;
- 7) the ICM is then factor analyzed (principal components algorithm) and
- 8) rotated to simple structure against a varimax criterion.

The resulting factors have been shown to correspond with major content themes in the document under analysis. Cluster algorithms have been applied successfully as well.

Whereas the empirical approach tends to have advantages over the dictionary approaches with respect to neutrality, the problems are here how to interpret the results of the factor or cluster analysis. Furthermore, problems arise from the fact that any empirical procedure is plagued by how many variables to incorporate. Certain approaches become inapplicable if the number of variables exceeds the number of cases. In content analysis, the number of documents or sections of documents is likely to be much smaller than the number of different words. Thus some selection

<sup>9</sup> Iker, H. P., Harway, N. J., A Computer Systems Approach Toward the Recognition and Analysis of Content, in: Gerbner, G., et al., *The Analysis of Communication Content*, New York 1969.

<sup>10</sup> Iker, op. cit.

is sensible. Selection can easily be oriented towards topical words, stylistic words, or whatever, but selection becomes a crucial determinant of the resulting description.

Iker has proposed some interesting empirical criteria for selecting words that enter into an empirical procedure. He correlates all terms with each other and then sums the 5th power of the correlation for each term, selecting these terms that have the largest sums. He selects the terms that have a subset of high sums of correlations rather than those terms that have many small inter-term correlations (SELECT procedure).

### 2.2.3. A Contextual Approach to Content Analysis

All methods discussed so far operate on the unit word. Contextual information is incorporated only modestly. This is where Cleveland, McTavish and Pirro come in with their QUESTER system<sup>11</sup>. Since the communication content is permanently changing, a communication model that attempts to analyze the communication process should be dynamic as well and should pay attention to the context. Words in a dictionary do not have a natural context, whereas words in communication do (exception: the HARVARD IV DICTIONARY contains disambiguation context). In addition words in a conversation do have known properties like syntactical structure, conceptual structure and contextual structure. These structures are interdependent, and like certain syntactical structures define the structures following (e. g. S-P-O), certain concepts are activated by preceding concepts. According to Quillian, a model should define which concepts are called by the nets of words in the context of a particular word. The relevant context defines these nets. QUESTER uses distance measures to define net boundaries.

### 2.2.4. Coding Approaches

Particularly for historical applications Couturier and Abehassera have developed the FORCOD system. Trained coders report in a standardized terminology on a tape recorder while reading the source text. The contents are represented in pairs of 'definers' and 'descriptors' (e. g. definer: occupation, possible descriptors: merchant, wine-merchant, wine-grower, carpenter). Then codes are assigned arbitrarily by the program to these descriptors. They may be recoded by intervention of the researcher to meaningful codes. These codes are analysed by the table analysis program FORTAB.

<sup>11</sup> Cleveland, E., et al., QUESTER: Contextual Analysis Methodology. Paper read at the ISSC Workshop on Content Analysis in the Social Sciences, Pisa 1974.



### 2.2.5. Retrieval according to subject for closer inspection

Over the last ten years powerful retrieval systems have been developed, like GOLEM and STAIRS, which are products of hardware manufacturers, RIQS, TEXT and Z.A.R., which have been specifically developed in the social science community<sup>12</sup>. According to a request by subject, they identify and retrieve from a document pool those documents which address the same topic. This function is available in some content analysis systems too (e. g. GENERAL INQUIRER). They may be misused to count how many documents were addressing a particular topic.

In particular, since many administrations have started to store information in computer accessible information pools these systems gain increasing importance. Prohibitively high costs may prevent their application for content analysis purposes on a larger scale even though the necessary routines are available in their software.

## 3. How Can Content Analysis Be Applied to Historical and Process-Produced Data?

The recent QUANTUM Documentation on Quantitative History 1977 lists several projects employing content analysis<sup>13</sup>, among these:

- Social structure of NSDAP. Analysis of its elites (Kalusche).
- Analysis of Prussian school books under the aspects 'education and industrialisation' (Lundgreen).
- Quantitative analysis of SA-elites social structure (Jamin).
- Abitur 1917–1971. Content analysis of graduation compositions (Mohler).
- Resistance in National Socialist Germany (Mann).
- Interest conflicts in trade politics during the German Revolution in 1848/49 (Best).
- Social status of candidates for the Reichstag 1898–1912 (Schroeder).
- Text analysis of Middle Latin chronicles (Arnold).
- Social Protest in 19th century Germany (Tilly).
- Marriage and family in German bishops letters to their flocks (Schmaelzle).
- The rise of a new elite: Social structure and political function of provincial administrators in France 1787–1820 (Reichardt).
- Content analysis of wills from 1648 to 1791 (Vogler).

<sup>12</sup> These systems operate on full text natural language. The retrieval process can be viewed as a reversed indexing (coding) process. The number of documents retrieved is displayed for each retrieval query. This could be used for content analysis purposes.

<sup>13</sup> Bick, W., et al., Quantitative historische Forschung 1977, Stuttgart 1977.

Those projects can be distinguished according to two goals. Most projects are analyzing phenomena for which statistical information is lacking. So they had to derive it from descriptive sources by a coding process. A minor number of projects is concerned with the orientation of the text source itself. For the time being the coding approach seems to be prevailing. This may be explained by the objective of revealing the social structure of groups primarily involved in historical events. The analysis of values or underlying intentions of the communicator has not yet been often focused upon<sup>14</sup>.

Some of the historical documents referenced in the QUANTUM documentation date back to the 12th century. The average time range for the period under investigation was calculated to be 114 years<sup>15</sup>. These time ranges put even more emphasis on the requirement to carefully consider relevant contexts. For historical documents the following should be considered as particularly relevant contexts<sup>16</sup>.

	Time (situational) context
Implicit context	Space (physical) environment
Total context	Linguistic (verbal) context
Explicit context	Extralinguistic context (kinesics)

Since the computer programs are designed to analyze the linguistic context, the researcher himself has to control the impact of the other contexts. Given that natural languages are a dynamically developing code (Latin may have been the only exception in the Middle Ages) we have to pay attention to changes in vocabulary and shifts in meaning over time. When investigating more recent documents, the researcher will be aware of socially redefined connotations. The latest significant example in the Federal Republic of Germany is the mention of 'Kreuth'. Before the decisive meeting of the CSU it was just a name of a village — at most important from a tourist's point of view. Thereafter it became synonymous with a dramatic discussion about the split between CDU and CSU.

We have to control these effects in order to avoid possible serious errors. This may imply thorough analysis of major events in a given period of time in addition to the documents under investigation. On the other hand we may draw inferences on unexpected changes of significance in word associations. Maybe they are clues to unknown social processes in the past for which no empirical evidence is yet available.

<sup>14</sup> Namenwirth, J. Z., Some Long- and Short-term Trends in One American Political Value: A Computer Analysis of Concern with Wealth in 62 Party Platforms, in: Gerbner, Analysis, pp. 223–241.

<sup>15</sup> Bick, Forschung, p. 12.

<sup>16</sup> Cf. Slama-Cazacu, T. S., Die dynamisch-kontextuelle Methode in der Sprachsoziologie, in: Kjolseth, R., and Sack, F., Zur Soziologie der Sprache, KZfSS, Special Issue 15 (1971), p. 82.

Analysis of Content and Context of Historical Documents —  
The Case of Petitions to the Frankfurt National Assembly 1848/49<sup>1</sup>

1. Document Analysis and Historical Social Research

If we study behavior, opinions or values in past societies, our data bases usually consist of written texts. In order to analyze such material historians developed the scrupulous techniques of 'source criticism', a research instrument which is successfully used to verify the reliability of texts, to reconstruct the course of political action on the level of cabinets and diplomatic services, or to write the history of ideas. The hermeneutical approach must fail if we want to analyze larger bodies of documents in a standardized way which is necessary if we are going to measure mass phenomena. In such cases we have to apply research instruments whose selective and evaluating procedures approximately correspond to the standards of survey research. The different approaches to content analysis which have been developed in the past decades, now offer a variety of tools to those historians and sociologists who want to study collective behavior, opinions or values of the past on the basis of written texts.

Nevertheless in Germany and moreover in Europe we can only find a few examples of the application of content analysis to historical source material. One reason may be the relatively large amount of work required by most approaches to content analysis. Furthermore, the primarily psychological orientation of some computer programs for text analysis does not correspond to the field of interest of many potential users. Another reason may be that different uses of language in past and present are not considered in the standard dictionaries. Finally we can observe that content

<sup>1</sup> The results presented in this paper are part of a more comprehensive study of the petitions to the Frankfurt National Assembly. For a more detailed description of methodological approaches and results, see Best, Heinrich, *Die quantitative Analyse inhaltlicher und kontextueller Merkmale historischer Dokumente. Das Beispiel der handelspolitischen Petitionen an die Frankfurter Nationalversammlung*, in: Best, Heinrich, and Mann, Reinhard (eds.), *Quantitative Methoden in der historisch-sozialwissenschaftlichen Forschung* (HSF, Vol. 3), Stuttgart 1977, pp. 162–205, and Best, Heinrich, *Interessenpolitik und nationale Integration 1848/49 – Handelspolitische Konflikte im frühindustriellen Deutschland*, Göttingen 1979. The research was partly financed by the Fritz Thyssen Stiftung.

analysis is rarely complemented by „contextual analysis“ of the documents in question: The examination of the political and social constellations which led to the formulation of a text, the institutional framework of its distribution and even such simple questions as who the authors were, and who the followers of its message were are often neglected. As opposed to this, the drafting and distribution of texts should be seen as a social process which involves a variety of individuals, groups or formal organizations. We can assume that the inclusion of such aspects in the repertory of content analysis will enrich our knowledge of both, text and context.

In the pages that follow I wish to give a short account of the methodological approaches and statistical techniques applied for coding and analyzing petitions to the Frankfurt National Assembly. My intention is to show that the combination of textual and contextual analysis can be used as a research technique which may help to discover motivational and structural determinants of political participation in past societies.

## 2. Relevance and Validity of the Source Material

The petitions to the Frankfurt National Assembly are important data bases concerning collective political behavior and public opinion in the the years 1848/49, as they were the most important institutional channel of mass political mobilization after the March Revolution. In the altered institutional framework the petitions enabled previously unpoliticized classes to involve themselves directly in political arguments and conflicts of interest<sup>2</sup>.

Roughly 25,000 to 30,000 petitions, articulating the demands on the first German parliament of about 2.5 to 3 million signers, disprove the thesis of „complete indifference of the mass and its passive role in public life“ after the March Revolution<sup>3</sup>. It is no surprise that the political, clerical and economic associations and pressure groups that had come into being in 1848 soon used the petitions as a manifestation and plebiscitarian justification of their demands. As there was also a great number of local and regional interest organizations, political corporations and informal groups included in the mobilization campaign, information concerning organizational patterns and conditions of the social structure of political action in early industrial Germany is to be expected from an analysis of the petitions.

<sup>2</sup> Literature on the petition movement is sparse and scattered. A broader but methodological unsatisfying study was written by Schirrmeister, Karl-Günther, *Menschenrechte in den Petitionen an die deutsche Nationalversammlung*, Phil. Diss. Bamberg 1970.

<sup>3</sup> See Hamerow, Theodore S., *The Elections to the Frankfurt Parliament*, in: *The Journal of Modern History*, 33 (1961), p. 29.

Main subjects of the petition movement were: basic human rights, inclusion of Austria in a national German state, demands for restrictive measures of trade regulation, for the separation of church and state, for abolition of feudal charges and for recognition of the German Imperial Constitution by the constituent states. One of the largest of these thematically quite precisely differentiated blocs is constituted by petitions relating to commercial policy (3,775 single petitions and 397,000 signatures). This category included all those petitions that made concrete demands concerning tariff policy, commented on the commercial policy of the National Assembly and/or the provisional central government of the Empire, or quite generally demanded the establishment of a German tariff union. Their relatively high share in the total amount of the petition movement (13 %–16 %) makes evident that violent conflicts of interest concerning commercial policy, which had been carried out during the period before the March Revolution (Vormärz) within the governments and parliamentary representations, on the diplomatic level and in the press, became the impulse of a mass movement under the different political conditions of 1848/49.

The significance of tariff policy as a means of economic control by the government as well as the relevance of commercial disputes for the exposure of cleavage structures are underlined by the fact that from the 1820s tariff conflicts caused serious political splits and dominated public debates not only in Germany but also in France, the United States and Great Britain<sup>4</sup>. Thus petitions relating to commercial policy make up a group of sources the analysis of which is likely to give information about mobilization processes, participation behavior and cleavage structures in early industrial Germany. Its special value lies in the possible combination of criteria of contents and context, thus allowing attribution of ideological and political tendencies to regional, social and organizational connections.

In order to limit the amount of potential research issues this paper will confine itself to the study of relations between the variations of the subject matters of the petitions and the social setup of the respective groups of signers.

Above all, the following questions are to be addressed:

- What were the social bases of recruitment for the protectionist and the free trade movement?
- What were the demands the signing socio-economic interest groups committed themselves to?
- What kinds of groups cooperated in the signing of petitions?
- Which factors influenced the cooperation of groups within the petition movement?

<sup>4</sup> See, e. g., Pincus, Jonathan J., *Pressure Groups and Politics in Antebellum Tariffs*, New York 1977; McCord, Norman, *The Anti-Corn Law League 1838–1846*, London 1958; Aydelotte, William O., *The Distintegration of the Conservative Party in the 1840s: A Study of Political Attitudes*, in: Aydelotte, William O., et al. (eds.), *The Dimensions of Quantitative Research in History*, Princeton 1972, pp. 319–346.

The findings are intended to contribute to the investigation of the widespread theory that class conflicts which were increasing during the revolution had prevented joint political action of bourgeoisie, farmers, petty bourgeoisie and workers, and had driven the bourgeoisie „into the arms of reaction“<sup>5</sup>.

### 3. Methods of Data-gathering

But before I begin to deal with the four research topics mentioned above, a short description of the concrete procedures of transforming „sources“ into „data“ is in order. In doing so I shall confine myself to the three selected variables, „tendency“, „reference“ and „signing interest group“ which are to be subjected to several methods of analysis in the next step.

The data have been collected using traditional interview procedures by means of a partly standardized questionnaire. Thus unexpected variants could be taken into consideration and the final definition of the variables could be taken up at the end of the data record.

The following data and groups of data for each petition were recorded: location information, date of composition, reception date, place of origin, affiliation to a collective petition, supporting organization, signing socio-economic pressure groups, number of signatures, and the references. Besides, a commentary is provided to record peculiarities. It was not until the completion of the partly standardized data inquiry that the variables were operationally defined and the final code plan fixed.

All of the petitions relating to trade policy could be included, as far as their contents is concerned, in one of the following categories: either „protectionist“, „free trade“ or „indifferent“. This classification was facilitated by the emergence in 1848/49 of commercial pressure groups such as the „General German Union for the Protection of National Labour“ (Allgemeiner Deutscher Verein zum Schutze der vaterländischen Arbeit) on the protectionist side, the „German Union for Freedom of Trade“ (Deutscher Verein für Handelsfreiheit) and the „Merchants' Union“ (Verein von Abgeordneten des Handelsstandes) on the free trader side. These organizations coordinated the great number of diffuse and sometimes even contradictory de-

<sup>5</sup> Today, this view is shared by marxist as well as ‚bourgeois‘ authors. See, e. g., Schmidt, Walter, Zur Rolle der Bourgeoisie in den bürgerlichen Revolutionen von 1789 und 1848, in: Zeitschrift für Geschichtswissenschaft, 21(1973), pp. 301–320; Dorpalen, Andreas, Die Revolution von 1848 in der Geschichtsschreibung der DDR, in: Historische Zeitschrift, 210 (1970), pp. 324–368; and Nipperdey, Thomas, Kritik oder Objektivität? Zur Beurteilung der Revolution 1848, in: Klötzer, Wolfgang, et al. (eds.), Ideen und Strukturen der deutschen Revolution 1848, Frankfurt 1974, pp. 143–162.

mands relating to tariff policy. Their tariff drafts fixed the „official“ sets of demands made by free trade and protectionist movements. Agreement or non-agreement with these competitive programs relating to commercial policy could generally be taken as a criterion for the inclusion of the petitions in the categories „protectionist“ or „free trade“.

Those petitions without specific demands relating to tariff policy, demanding merely the establishment of economic unity throughout Germany, reconciliation of antagonistic points of view or consultation of experts for the formulation of the tariff, were labelled „indifferent“.

Petitions relating to trade policy distributed among the three basic categories of contents as follows:

*Tab. 1: Distribution of Petitions Among Basic Categories of Contents*

tendency	absolute frequency	relative frequency (in percent)
protectionist	3,400	90.1
free trade	347	9.2
indifferent	28	0.7
	<u>3,775</u>	<u>100.0</u>

Apart from this rough classification of the petitions according to contents, the individual grievances were listed as well. It soon turned out that the sets of grievances were relatively deficient in variants; more than 89 % of the petitions belonged to a total of 52 collective petitions the texts of which had been previously formulated and supraregionally distributed. Central control of the agitation by free trade and protectionist leagues contributed to further standardization of contents.

These conditions enabled us to do without a complicated coding pattern: The grievances we identified were defined as dichotomous variables, which means that each petition was classified according to presence or absence of every individual reference. Reasons and commentaries given in the petitions were not embodied in the data set, but were nevertheless considered significant material and recorded in the questionnaires. A selection of the grievances recorded is to be found in Table 4.

2,772 petitions (= 73.4 % of all petitions relating to trade policy) contained information about the social position (Stand) of the signers, appearing sometimes as a collective label (example: „We, the Tritenheim wine growers, demand . . .“), from November 1848 on as an addition to the individual signature. An allusion to a heterogeneous group of supporters was an important argument in favour of commercial demands, and it became a common practice to provide one column in the list of signatures for the signers to enter their „social position“ (Stand).

With regard to the extremely diffuse statements characterizing „social position“ overall social categories had to be established to take into account peculiarities of material as well as special problems of research. Given these conditions, the formation of pressure groups organization in 1848 seemed to be the most appropriate pattern for classification of the various descriptions of occupation; it represented an immediate clue to problems relating to policy of interest and at the same time took into consideration the contemporary perception of the social system.

In contrast, I refrained from imposing a model of social stratification or categories of occupation lacking an equivalent in contemporary conceptions, as it can be assumed that specific forms of political action and collective ideological tendencies have been developed in the actual context of cooperating pressure groups rather than within the framework of social categories defined *ex post facto*. This way I was also able to avoid certain problems of stratification and grouping that tend to obstruct analysis of social phenomena especially of the 19th century.

The definition of the socio-economic pressure groups used for social classification thus follows the pattern for organization of economic pressure groups in 1848/49. Farmers organized in a „Congress of Delegates of German Agricultural Associations“ (Kongreß der Abgeordneten deutscher landwirtschaftlicher Vereine), master craftsmen in the „German Congress of Craftsmasters“ (Deutscher Handwerker- und Gewerbekongreß), workers in the „General German Worker's Congress“ (Allgemeiner deutscher Arbeiterkongreß) and in the „Berlin Workers' Congress“ (Berliner Arbeiterkongreß), merchants in a „Merchants' Union“ (Verein von Abgeordneten des Handelsstandes) and manufacturers in the „General German Union for the Protection of National Labour“ (Allgemeiner Deutscher Verein zum Schutz der vaterländischen Arbeit), which had been founded as a manufacturers' organization. An extra group such as the wine growers articulated their demands in the „Assemblies of German Wine- and Fruit Producers“ (Versammlungen der Deutschen Wein- und Obstproduzenten), employees in the textile industries also founded their first organization, the „Association of Saxon Spinning Clerks“ (Verein Sächsischer Spinnereibeamter). These national assemblies and top organizations were established on a broad basis of local organizations and informal groups; one can thus infer that many everyday social contacts took place within the framework of socio-economic interest groups as well.

Apart from the socio-economic pressure groups mentioned here were, however, some groups that had made no effort to form organizations, although they had been assigned a clearly limited position in society. Either motivational or material resources for the development of a corporate structure were lacking in these cases. This applies in the first place to civil servants, to members of academic professions, and to the largest group in early industrial society – the agricultural labourers. Because of their special significance as supporters or subjects of the trade mobilization campaign these groups were included in the set of variables as independent social categories in spite of the lack of specific corporations. With regard to these exceptions I have assigned the given descriptions of occupation to the ten socio-economic pressure groups: manufacturers, merchants, farmers, master craftsmen, work-



ers, clerks, wine- and tobacco growers, members of academic professions, civil servants and agricultural labourers. The inclusion of individual professions in these groups tended to follow the contemporary craft statistics which was to a certain extent divided into similar categories<sup>6</sup>.

Recording the data, I defined each pressure group as a dichotomous variable and registered its presence or absence in the lists of signatures. Thus the affiliation of an individual signature to a certain group was not recorded. Consequently we cannot make statements like „in petition X, 10 % of the signers were manufacturers“ but only statements such as „workers signed in 20,3 % of all petitions“. This procedure had its basis and justification in the hypothesis that the signing of petitions was the result of joint decisions made within the framework of socio-economic pressure groups and, thus, represented the intention of the group.

This assumption is supported by the fact that in 62,5 % of the cases in which information about the signers' social position was available the respective pressure groups signed petitions isolated from each other. A further 27,5 % of cases concentrated on 20 types of group cooperation in signing petitions (example: manufacturers together with workers and clerks). On the whole, 90 % of the petitions were distributed among 30 types of group representation which make up 2,9 % of a total of 1,024 theoretical possibilities of cooperation. This finding signifies that the members of the various pressure groups signed either on their own or selectively together with an exclusive set of social partners — a finding that definitely supports the initial hypothesis.

#### 4. A Relational Analysis of Content and Context of Petitions to the Frankfurt National Assembly

Subsequent to this short description of the set of variables that was analyzed I am going to answer the four research questions referred to above.

In 1848/49, when plebiscitary support of commercial demands was supposed to have an effective bearing on political decisions, mobilization of a numerous and socially heterogenous following was more than ever the aim and measure of success in agitation by protectionists as well as free traders. In this stage the different bases of recruitment of protectionist and free trade petition movements show the possibilities as well as the limits of their wooing adherents.

<sup>6</sup> See, e. g., *Statistische Mitteilungen aus dem Königreich Sachsen, Dresden 1854*, and *Tabel-  
len und amtliche Nachrichten aus dem preußischen Staat für das Jahr 1849*, Vols. 4–6, Berlin  
(no year).

**Tab. 2: The Representation of Socio-Economic Pressure Groups in the Protectionist- and Free Trade Movements**

pressure groups	protectionist petitions		free trade petitions	
	N	% *	N	% *
manufacturers	411	16.0	15	6.1
merchants	300	11.9	168	68.3
farmers	523	20.8	74	30.1
master craftsmen	1,094	43.6	51	20.7
workers	753	30.0	12	4.9
clerks	81	3.2	5	2.0
public servants	74	2.9	22	8.9
academic professions	27	1.1	16	6.5
agricultural labourers	156	6.2	38	15.4
wine- and tobacco growers	852	33.9	0	0.0
	N = 2,511		N = 246	

\* percentages are adding up to more than 100 %.

**Tab. 3: Distribution of Pressure Groups between the Protectionist and Free Trade Movements**

pressure groups	protectionist petitions				free trade petitions			
	N	N	%*	dev.from** mean	N	%	dev. from*** mean	
manufacturers	426	411	96.5	+ 5.4	15	3.5	- 5.4	
merchants	468	300	64.1	- 27.0	168	35.9	+ 27.0	
farmers	597	523	87.6	- 3.5	74	12.4	+ 3.5	
master craftsmen	1,145	1,094	95.5	+ 4.4	51	4.5	- 4.5	
workers	765	753	98.4	+ 7.3	12	1.6	- 7.3	
clerks	86	81	94.2	+ 3.1	5	5.8	- 3.1	
public servants	96	74	77.1	- 14.0	22	22.9	+ 14.0	
academic professions	43	27	62.8	- 28.3	16	37.2	+ 28.3	
agricultural labourers	194	146	80.4	- 10.7	38	19.6	+ 10.7	
wine- and tobacco grow.	852	852	100.0	+ 8.9	0	0.0	- 8.9	

\* row-values of columns 3 and 6 are adding up to 100.0%.

\*\* mean share of protectionist petitions in all petitions concerning tariff policy: 91.1 %.

\*\*\* mean share of free trade petitions in all petitions concerning tariff policy: 8.9 %.

In interpreting the results of Table 2, the different degrees of success of free trade and protectionist movements must, of course, be taken into account: There were 2,511 protectionist petitions supplying information about the signing pressure groups compared to no more than 246 corresponding petitions for free trade. For example farmers were represented in 30 % of all free trade petitions, whereas 87.6% of all petitions in which farmers were represented, were protectionist (see Table 3).

Tables 2 and 3 reveal that free traders remained confined to a great extent to their traditional core of adherents — the merchants. Traditions of the period before the March Revolution were resumed by rather frequent representation of civil servants and members of academic professions. The groups of farmers and agricultural labourers which were likely to be in favour of free trade at least in those cultivation areas depending on export, were represented above average in the free trade petitions; but nevertheless the protectionist advantage could not nearly be made up for even in the agricultural sphere. Groups such as manufacturers, master craftsmen, workers and clerks were represented below average.

On the other hand, the protectionists found the main part of their following among the industrial and agricultural middle- and lower classes. The group of wine and tobacco growers, which was not at all represented in free trade petitions, held the second place of all categories. On the whole, members of agricultural professions were represented in 55.5 % of all protectionist petitions, workers and master craftsmen in 57.7 %. In view of the success in mobilizing these groups, the traditional core of adherents of the protectionist movement — the manufacturers — lost some of its import. Nevertheless manufacturers committed themselves to protectionist petitions to a considerably greater extent than merchants committed themselves to free trade petitions. It comes as a surprise that this protectionist advantage even applied to the group of merchants that voted much more frequently for restrictions by means of tariff policy than for free trade.

One reason for the protectionist movement's vast success in mobilizing the various socio-economic pressure groups might have been their disposal to match the uniform and purposeful agitation of the „General German Union for the Protection of National Labour“ (Allgemeiner deutscher Verein zum Schutze der vaterländischen Arbeit). This advantage, though, is but an insufficient explanation for the protectionist success as the propaganda impulses had to go along with appropriate motivational incentives and readiness for cooperative political action on the part of the pressure groups in order to lead to the signing of trade petitions: It is also the different reactions to agitation that may serve as an explanation for the different mobilization effects as well as different bases of recruitment of protectionist and free trade movements.

If one wants to reveal the motives of pressure groups for their participation in the trade petition movement, one could reasonably take it as an initial hypothesis that they signed petitions above all because of their commitment to individual topics contained in the petitions. In fact no arbitrary relations between readiness to sign petitions and the existence of certain subject matters are to be expected: pressure groups frequently voted for specific sets of demands submitted to them either

by their own organizations or by national free trade or protectionist associations, the aim being the achievement of the highest degree of mobilization that could possibly be reached.

I have depicted the relations between the variables „signing pressure group“ and „reference“ by means of a matrix of  $\Phi$  coefficients. In the present case,  $\Phi$  reaches its maximum value if the appearance of a group always goes along with the occurrence of a certain subject matter, whereas other groups do not vote for this grievance. In case a group tended not to appear in the lists of signatures referring to a special subject matter, the value is negative<sup>7</sup>. As it can hardly be decided whether this abstention is due to indifference or deliberate refusal on the part of the group not represented, these negative coefficients have been neglected. In order to facilitate the interpretation of the results, it was decided that the same should apply to coefficients below .10 (rounded up). If measuring the relation between the occurrence of a certain subject matter and the signing of the respective petitions by a pressure group resulted in a coefficient of  $\Phi \geq .10$  this was regarded as evidence for the ‚commitment‘ of this group in favour of the subject matter in question.

Detailed interpretation of all the results of Table 4 would go beyond the scope of this paper. Nonetheless I would like to record three points that seem important to me:

- Obviously, there was only a small number of demands distinctly appealing to specific groups (wine growers – wine tariffs etc.). Most of the grievances were supported by a heterogenous following.
- Common attachment to certain grievances united different pressure groups. It must be considered whether this common articulation of interests led to direct interaction of the groups within the petition movement.
- Individual groups committed themselves to demands that were directly opposed to each other. At this point the effect of a distinctly sectional cleavage structure becomes obvious, splitting up for instance the merchants in their entirety into two groups: one group of merchants from northern Germany, committed to export trade and clearly preferring free trade measures; and a second group of merchants from the western and southern parts of Germany, committed to inland trade and protective tariffs.

<sup>7</sup> The version of  $\Phi$  used in this paper (with signs) is described in Benninghaus, H., *Deskriptive Statistik*, Stuttgart 1977. The calculation of  $\Phi$  was derived from 2 x 2 tables of the following type:

		protest against free trade tariff		
		not mentioned	mentioned	
master craftsmen	not represented	566	987	1553
	represented	543	512	1055
		1109	1499	2608
$\Phi = 0.15$				

Tab. 4: „Commitment“ of Pressure Groups in Favour of Particular Subject Matters  
( $\Phi$ coefficients)

Grievances*	Manuf.	Merch.	Farm.	Mast.	Work.	Vine grow.
Support of an immediate establishment of the German commercial union	0.18	0.12		0.17	0.10	
Support of the convocation of a congress of experts in order to decide the tariff questions	0.13		0.12	0.11	0.15	
Support of raising tariffs on iron	0.10				0.26	
Support of raising tariffs on textiles	0.10			0.10		
Support of decreasing tariffs on raw materials and foodstuff				0.22		
Refusal of export duties				0.20		
Support of export duties	0.13					
Support of a combination of protective tariffs and restrictive guild-laws				0.15		
Support of high tariffs on wine						0.92
Support of high tariffs on tobacco						0.12
Support of high tariffs on sugar			0.15			
Support of reciprocal trade agreements	0.21	0.14	0.13	0.35	0.28	
Refusal of reciprocal trade agreements		0.31				
Support of the motion of Eisenstuck and others	0.17					
Refusal of the motion of Eisenstuck and others		0.15				
Support of the draft of a protectionist tariff	0.10	0.17		0.29		
Support of the general aims of the protectionist league			0.20			0.10
Support of the draft of a free trade tariff		0.34				

\* Only such grievances were noted, to which the commitment of at least one pressure group was indicated by a  $\Phi \geq 0.10$ . The commitment of public servants to particular grievances was generally lesser than the liminal value of  $\Phi = 0.10$ . Holders of academic professions tended to reject the motion of Bernhard Eisenstuck ( $\Phi = 0.13$ ) and were inclined to support the free trade tariff ( $\Phi = 0.10$ ). Agricultural labourers tended to support the general aims of the protectionist league ( $\Phi = 0.15$ ). On the other hand they were inclined to support the free trade tariff ( $\Phi = 0.13$ ). The texts of some grievances had to be abbreviated in case of space shortage.

Tab. 4: „Commitment“ of Pressure Groups in Favour of Particular Subject Matters  
( $\Phi$ coefficients)

Refusal of the draft of a free trade tariff					0.58
Distrust of the Frankfurt National Assembly in case of its protectionist prepossessions	0.15				
Protest against the retardation of the pro- tectionist program by the National Assembly and/or the minister of trade	0.15	0.18	0.25	0.50	

At this point the question arises as to what extent participation of the various socio-economic pressure groups in the petitions can be explained by commitment to the subject matters of the petitions. As opposed to the preceding analysis, we shall refrain from investigating how far the individual demand was connected with participation or non-participation of a pressure group in petitions. We shall rather examine how far participation was influenced by the set of grievances as a whole.

In order to ascertain the strength of this relationship,  $r^2$  was used, giving the „proportion of explained variation in per cent“. In the present case, e. g., the value of the (multiple) correlation between occurrence of certain grievances in the petitions and representation of craftsmen in the lists of signatures of  $R^2 = 0.51$ , was interpreted to the effect that participation or non-participation can be „explained“ by the presence or absence of certain grievances to a total of 51%<sup>8</sup>.

Tab. 5: „Commitment“ of Pressure-Groups to the Set of Grievances as a Whole

pressure groups	multiple $R^2$ : total set of grievances/representation of interest groups
manufacturers	$R^2 = 0.22$
merchants	$R^2 = 0.28$
farmers	$R^2 = 0.27$
master craftsmen	$R^2 = 0.51$
workers	$R^2 = 0.50$
clerks	$R^2 = 0.03$
public servants	$R^2 = 0.03$
academic professions	$R^2 = 0.05$
agricultural labourers	$R^2 = 0.12$
wine- and tobacco growers	$R^2 = 0.91$
	in all cases: $s = 0.000$

<sup>8</sup> The assumption that the grievances were independent variables was derived from the fact that in most cases signers supported previously formulated texts. Seldom they influenced the formulation of petitions directly. For the justification of an application of multiple correlation analysis to dichotomous variables see Blalock, Hubert M., Social Statistics, 2nd. ed., New York 1972, pp. 454 ff., 498 ff.

The results reveal characteristic differences between the pressure groups as regards their commitment to the sets of demands offered.

Clerks, public servants and members of academic professions tended but to a small degree to focus on certain subject matters of the petitions. Their indifference suggests that they did not develop any independent centers of interest as far as tariff policy is concerned, and that their participation in the petition movement was only a subsidiary one. The same applies to the agricultural labourers whose mobilization was evidently only slightly more connected to the presence of certain demands in petitions. This outcome remarkably corresponds with the fact that in 1848/49 the groups mentioned did not establish any independent organizations on the national level. In contrast to these groups, wine- and tobacco growers petitioned almost exclusively for grievances referring to their own specific interests. Petitions of workers and master craftsmen included subjects relating to specific groups as well, though without focusing on an exclusive subject. Obviously, however, specific demands had to be contained in the texts in order to induce them to participate.

On the other hand, manufacturers, merchants and farmers concentrated less distinctly on special grievances. In the case of the two groups mentioned first, their relative indifference can be accounted for by the fact that they were the initiators of the free trade and protectionist petition movements. So they were directly involved in the distribution of different text versions and in this function signed copies the demands of which were traced out for other groups as well. In the case of those farmers not specialized in wine- and tobacco growing, who, with the exception of sugar customs, did not make any demands for agrarian protectionism, relative indifference is rather due to a „hanger-on-effect“; an assumption that shall be supported by further evidence elsewhere.

To summarize, it is to be kept in mind that „issue-orientation“ was no doubt an important motive for signing trade petitions. With the exception of wine and tobacco growers, however, this factor by itself is only an insufficient explanation for participation of pressure groups in the petition movement.

An indication of further attempts at an explanation was given by the allusion to a „hanger-on-effect“: Obviously the social relations between pressure groups had a bearing on their participation in the protectionist and free trade petition movement. A comparison of the values given in Table 5 for commitment in favor of sets of grievances on the one hand and relative frequency of isolated signing of texts by pressure groups on the other hand supports the assumption that „issue-orientation“ and involvement in systems of social relationship were complementary impulses for participation of pressure groups in petitions.

The results reveal that groups which were distinctly attached to certain grievances, tended to sign petitions on their own, whereas groups which were relatively indifferent to the subject matters, petitioned almost exclusively together with their social partners. Table 6 supports the assumption that analysis of social relations established by pressure groups within the framework of the petition movement suggest further evidence as to the impulses for their participation. For analysis of the patterns of interaction reflected by the lists of signatures, I made also use of Configu-

Tab. 6: „Commitment“ of Pressure Groups in Favour of the Set of Grievances and Relative Frequencies of Isolated Signing of Petitions

pressure groups	„commitment“	frequencies of isolated signing	
		N	%*
manufacturers	$R^2 = 0.22$	109	25.2
merchants	$R^2 = 0.28$	150	31.8
farmers	$R^2 = 0.27$	78	13.0
master craftsmen	$R^2 = 0.51$	364	31.6
workers	$R^2 = 0.50$	203	26.5
clerks	$R^2 = 0.03$	5	5.8
public servants	$R^2 = 0.03$	3	3.1
academic professions	$R^2 = 0.05$	0	0.0
agricultural labourers	$R^2 = 0.12$	11	5.8
wine- and tobacco growers	$R^2 = 0.91$	809	94.8

\* The proportion is related to the total number of petitions signed by the respective pressure groups (basic values, see Tab. 3, col. 1).

ration Frequency analysis (CFA)<sup>9</sup>, a statistical instrument developed by E. A. Lienert and J. Krauth for use in psychology. CFA measures the frequency of occurrence of particular configurations — in the present case for instance joint signing of protectionist petitions by workers, manufacturers and clerks — and compares this value to the probability of occurrence of the respective configuration (expected value). Inherent structures of cooperation of the various pressure groups may be taken for granted if the frequency of occurrence of a certain pattern of interaction is significantly higher than expected. The configuration „manufacturer, worker, clerk“ for instance is to be expected only once, whereas there are in fact 22 protectionist petitions jointly signed by these groups; a discrepancy between expected value and observed value revealing an underlying regularity that requires interpretation.

The results of CFA have been visualized by a diagram using the technique of Smallest Space Analysis<sup>10</sup>. By this means the different pressure groups were arranged

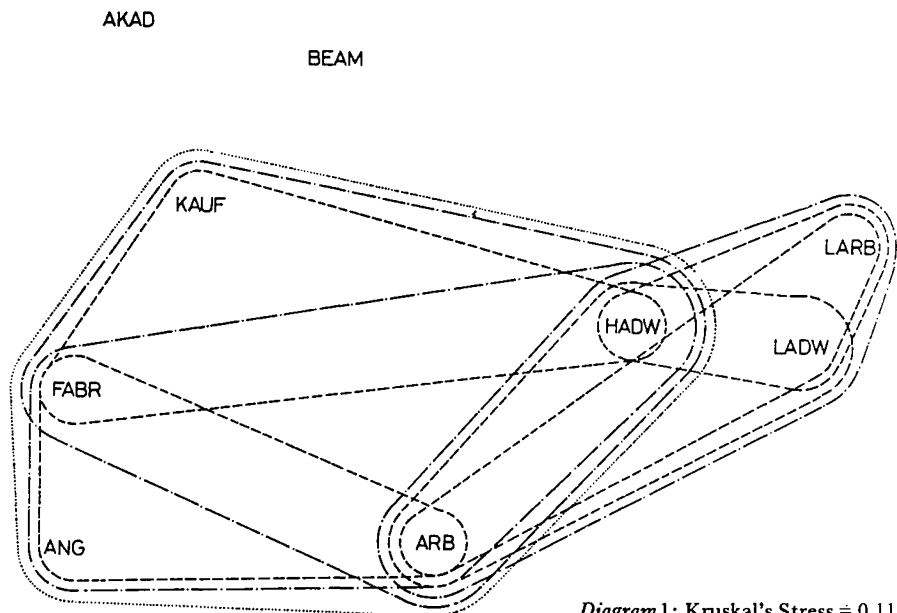
<sup>9</sup> See Krauth, Joachim and Lienert, Gustav Adolf, KFA. Die Konfigurationsfrequenzanalyse und ihre Anwendung in Psychologie und Medizin, Freiburg 1973. For computerized analysis I used the program KONFA which was written by Burkhard Roeder. For a more detailed description of KONFA see Roeder, Burkhard, Die Konfigurationsfrequenzanalyse nach Krauth und Lienert. Ein handliches Verfahren zur Verarbeitung sozialwissenschaftlicher Daten, demonstriert an einem Beispiel, in: Kölner Zeitschrift für Soziologie und Sozialpsychologie, 26 (1974), pp. 819–844.

<sup>10</sup> Lingoes, James C., The Guttman-Lingoes Nonmetric Program Series, Ann Arbor/Mich. 1973. For the Analysis of the Petitions I used the program SSA I/MINISSA I.



within a two-dimensional system of coordinates in such a way that groups cooperating rather frequently, such as farmers and agricultural labourers, are plotted close to one another, whereas groups which rarely signed petitions together, such as workers and academics, are plotted farer apart. Thus social distance as far as it was revealed during the petition movement is measured by the distance between the different groups, whereas the characteristic types of group cooperation, as identified by aid of CFA, are marked by the connecting lines (see Diagrams 1 and 2).

**Diagram 1: Interaction Pattern of the Protectionist Petition Movement**



*Diagram 1: Kruskal's Stress = 0.115*

**Legend to Diagrams 1 and 2:**

- = significant frequent isolated signing of one group
- ↔ = significant frequent interaction between two groups
- = significant frequent interaction between three groups
- ..... = significant frequent interaction between four groups
- ..... = significant frequent interaction between five groups

FABR = manufacturers  
 KAUF = merchants  
 HADW = master craftsmen  
 LADW = farmers  
 ARB = workers

LARB = agricultural labourers  
 ANG = clerks  
 AKAD = academic professions  
 BEAM = public servants

Diagram 2: Interaction Pattern of the Free Trade Petition Movement

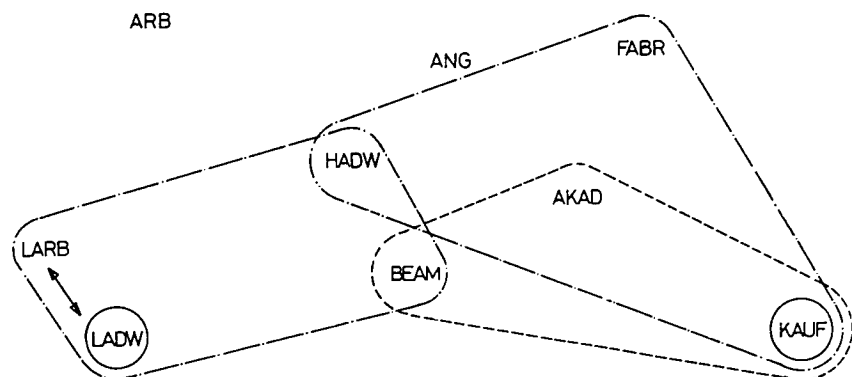


Diagram 2: Kruskal's Stress = 0.089

Two clearly different substructures can be discerned regarding the social interaction pattern of the protectionist petition movement: a „rural“ and an „urban“ network, master craftsmen and workers acting as intermediaries between them. Close relationship between workers, master craftsmen, farmers and agricultural labourers can be accounted for by the connection of proto-industry and agriculture within the framework of rural domestic industry. Obviously the protectionist movement succeeded in exploiting for its mobilization campaign the close social contacts developing between the different producer groups in the country. Because of their simultaneous affiliation to „rural“ and „urban“ milieu craftsmen and workers developed much more complex patterns of interaction than the other groups. As far as workers are concerned, this finding contradicts the obvious suspicion that they generally had been forced to participate in the protectionist movement by pressure exerted by the manufacturers: only 26.9 % of all protectionist petitions that were signed by workers were at the same time signed by manufacturers, whereas in 26.5 % of all cases workers petitioned without any participation of other groups. In the case of the master craftsmen we are struck by an especially close connection to the farmers, linking them closer to the rural than to the urban milieu.

Manufacturers belonged to two different fields of interaction as well. A system of relations, also including clerks and workers, can be accounted for by the cooperation of these groups within the framework of industrial companies, whereas joint signing of petitions with merchants and craftsmen reflects relations between industrial and commercial elements of urban economy. Academics and civil servants did not have any significant relations to other groups within the framework of the protectionist petition movement.

Within the free trade petition movement too, we have a „rural“ as well as an

„urban“ pattern of interaction connected by craftsmen and public servants<sup>11</sup>. Cooperation of the groups involved on the free trade side, however, was significantly rarer than on the protectionist side. It comes as a surprise that merchants, being the most important supporters of the free trade movement, held a peripheral position within its system of social relations. This gap between the core of the interested and the group of people addressed can be taken to be one reason for the failure of the free trade mobilization campaign<sup>12</sup>.

In the free trade movement groups concerned tended to cooperate less and less as time went on: the proportion of petitions with isolated group representation increased from 62,0 % of 78,2 % subsequent to November 1848 when the campaign in favor of the free trade tariff project was unleashed. Contrary to the intention of its initiators the free trade tariff had anything but a supporting and unifying effect for an overall free trade petition movement but rather seems to have increased the isolation of north German foreign trade merchants.

<sup>11</sup> To make possible a comparison between free trade petitions and protectionist petitions I eliminated those petitions which were supported by wine- and tobacco growers. This procedure did not affect the patterns of interaction of the other groups, because wine- and tobacco growers signed nearly in all cases without participation of other groups. Configurations were noted if their probability of occurrence was  $< 0.1\%$  and if their absolute frequency was  $\geq 5$ .

interacting pressure groups	expected value	frequency
FARM, WORK, AGRLABR	4	19
FARM, MAST, AGRLABR	7	41
FARM, MAST, WORK	37	108
FARM, MAST, WORK, AGRLABR	3	13
MANF, WORK, CLERK	1	22
MANF, MAST, WORK, CLERK	1	10
MANF, MERCH, MAST	13	37
MANF, MERCH, MAST, WORK	5	32
MANF, MERCH, MAST, CLERK	0	8

N = 2510

<sup>12</sup> In the case of free trade petitions I used the same selective procedures as for protectionist petitions. Only the liminal value for the frequency of configurations was reduced from 5 to 3, because the number of free trade petitions was significantly smaller than the number of protectionist petitions.

interacting pressure groups	expected value	frequency
FARM (isolated)	9	33
FARM, AGRLABR	7	12
FARM, MAST, PUBSERV, AGRLABR	0	3
MERCH (isolated)	7	124
MERCH, PUBSERV, ACADS	0	3
MANUF, MERCH, MAST, ACADS	0	3

N = 246

On the other hand, protectionist patterns of interaction gradually took shape: Prior to November 1848, which means before agitation was controlled by the protectionist league, 87,8 % of protectionist petitions subjected to CFA were signed by the pressure groups separately; subsequent to November 1848, this proportion went down to 28,3 %. Formation of the characteristic patterns of interaction, as for instance development of „rural“ and „urban“ systems of relations, did not occur until the second stage.

Reasons for the significant change of participation behaviour can be found in the different techniques of organization of petitions and in the different tenor of protectionist agitation. Prior to November 1848, organizations and grievances specifically concerning particular groups prevailed, whereas the mobilization campaign of the „General German Union for the Protection of National Labour“ increasingly relied on overall patterns of organization combining concessions to specific interests with an appeal to the common „material interests of the nation“. In spite of the fact that latent conflicts continued to exist the protectionist program displayed considerable integrating power. Only the wine- and tobacco growers were mobilized almost exclusively by specific demands and largely without participation of other groups.

Analysis of the patterns of interaction by means of CFA revealed that cooperation of groups was determined by their affiliation to a particular „milieu“. Remember the discovery of „rural“ and „urban“ systems of relations or the significance of working in the same factory. At this point the question arises which basic social conditions and which motives influenced group cooperation and, on the other hand, which weight has to be assigned to these determinants.

In order to reveal connections of this kind, factor analysis seems to be the most appropriate instrument. In the present study entries announcing presence or absence of the ten socio-economic pressure groups in the petitions were intercorrelated as original variables. Factors influencing their relations were interpreted as clustering forces determining cooperation of these groups within the petition movement. Free trade and protectionist petitions were analyzed separately as CFA had revealed different patterns of interaction, and comparison of the contrasting groups promised additional possibilities of interpretation.

As to the protectionist movement, factor analysis revealed four determinants of group cooperation<sup>13</sup>. One factor, identified as „working together“, most significantly influenced relations between pressure groups. It accounts for 50,3 % of variance and above all determined cooperation of manufacturers, workers and clerks. Second was one factor signifying „living together in the country“ (21,7 % of variance), showing high factor scores for farmers and agricultural labourers, and another factor denoting „perceived identity of interests“ (16,4 % of variance), showing high scores for craftsmen and lower scores for manufacturers, merchants, workers and farmers. A factor standing for „social proximity of local dignitaries“ was least

<sup>13</sup> I used SPSS, vers. CDC 6.0, subprogram FACTOR, method PA 2, orthogonal solution

significant (11.6 % of variance). This factor showed high scores for public servants and academic professions, and considerably lower scores for manufacturers and merchants.

Almost the same determinants of group cooperation are discernible within the free trade petition movement<sup>14</sup>. Only the factor denoting „social proximity“ is lacking here. Within the free trade petition movement the factor standing for „living together in the country“, accounting for 45.9 % of variance, was of the greatest importance for cooperation behaviour, succeeded by the factor signifying „perceived identity of interests“ with 41.5%. The factor identified as „working together“, which showed high scores for craftsmen and workers within the free trade movement, follows with 12.6 %.

(VARIMAX). See Nie, Norman H., et al., SPSS: Statistical Package for the Social Sciences, 2nd ed., New York 1975, pp. 480 ff.

The following factor scores were calculated:

	FACTOR 1 („working together“)	FACTOR 2 („living together“)	FACTOR 3 („perceiv. id. of interests“)	FACTOR 4 („social proximity“)
MANUF	0.56	— 0.15	0.19	0.14
MERCH	0.21	— 0.07	0.29	0.26
FARM	— 0.03	0.65	0.20	0.02
MAST	— 0.04	0.20	0.73	0.06
WORK	0.48	0.30	0.15	— 0.06
CLERK	0.46	0.00	0.04	0.10
PUBSERV	— 0.06	0.12	0.02	0.60
ACAD	0.07	0.02	0.04	0.52
AGRLABR	0.02	0.47	0.05	0.09
WINEGROW	— 0.41	— 0.34	— 0.71	0.02
PCT of Var.	50.3	21.7	16.4	11.4

<sup>14</sup> Factor scores:

	FACTOR 1 („living together“)	FACTOR 2 („perceiv. id. of interests“)	FACTOR 3 („working together“)
MANUF	— 0.05	0.35	— 0.02
MERCH	— 0.95	0.16	— 0.01
FARM	0.70	0.11	0.01
MAST	0.12	0.53	0.57
WORK	0.08	0.04	0.61
CLERK	0.04	0.37	0.13
PUBSERV	0.05	0.66	0.09
ACAD	— 0.10	0.66	0.04
AGRLABR	0.55	— 0.05	0.22
PCT of Var.	45.9	41.5	12.6

The results of factor analysis suggest that existing systems of social relations like „living together“ and „working together“ played a more important role for cooperation of pressure groups within the petition movement than mere ideological determinants such as the factor standing for „perceived identity of interests“ did. The protectionists were able to benefit from existing social relations within social sub-systems to a greater extent than free traders, whose more abstract grievances led to a more ideologically determined cooperation of pressure groups.

## 5. Conclusion

The results presented have revealed that the assumption that class conflicts had increased during 1848 and 1849 and prevented joint political action of bourgeoisie, workers, farmers and craftsmen, is at least not true of the trade petition movement. Especially the protectionist movement actually succeeded in expanding the basis of recruitment for their petition campaign beyond the limits of status and milieu. Conflicts between socio-economic pressure groups could be concealed by nationalist economic propaganda at least for the topic under consideration and at least during the period under research.

It was only subsequent to November 1848, in a period when according to most authors, the bourgeoisie fled „into the arms of reaction“ for fear of the social demands of the proletariat and petty bourgeoisie, that the patterns of interaction of the groups participating in the petition campaign took shape and the protectionist mass movement reached its climax

The need to revise some current views of the Revolution of 1848/49 – which I experienced myself in the necessity of revising some of my own expectations – indicates once more that knowledge in collective behavior and opinions of past societies cannot be derived from scattered citations or with reference of contemporary authorities. As opposed to that it needs gathering of mass data in a standardized way and the application of statistical analyses. Content analysis will be an important research instrument for a „collective history“ of this type, whose databases nearly exclusively consists of written text. Systematic inclusion of a „contextual analysis“ into the repertory of content analysis could be essential to its adoption and expansion in the field of quantitative history because contextual analysis of historical documents opens further dimensions of research and could form a methodological link to traditional historical „source criticism“.



## V. Time Series Analysis





## Procedures for Periodizing History: Determining Distinct Eras in the Histories of Britain, France, Germany and Italy\*

Fernand Braudel in the prefaces to the editions of his great work *The Mediterranean World in the Age of Philipp II* emphasizes the contradiction at the heart of all efforts to study societies over the *longue durée* while still acknowledging the force and pull of *l'histoire événementielle*. Braudel responds to this conflict by finding event centered history dangerous and misleading. Historians and social scientists indebted to his work and to that of the *Annales* school have imitated his decision to give priority to the *longue durée*. Problems persist, however, even after this theoretical and methodological choice has been made. How shall the periods that identify the essential continuities and discontinuities in a society be established? What procedures most effectively establish the periodization critical to the analysis of a society? What methods make it possible to test the validity of the periodization adopted in historical and sociological analysis? Why does it matter? What are the substantive issues at stake for history and the social sciences?

Answers to these questions must first take into account methodological and theoretical problems of measurement in the construction of time series data. Secondly, it is necessary to examine the possibilities and limits of periodizing on the basis of a single variable. Thirdly, and only after the other issues have been considered, it is necessary to consider the implications of periodizations that are possible employing multiple variables.

In the sections that follow some light will be thrown on these issues by using, for illustrative purposes, the problem, of periodizing and analyzing the growth of public expenditures for education, health, welfare, and social security in Britain, France, Germany, and Italy from 1870 through 1965. In the last three decades of the nineteenth century and in the twentieth century expenditure for these purposes was at the nexus of the changes affecting the modernization of these societies.

\* Prepared with the support of the National Science Foundation (SOC 75-192821), the Ford Foundation, the Graduate Research Committee, University of Wisconsin, and the Institute for Research on Poverty, University of Wisconsin. Their financial support does not indicate acceptance of the ideas or opinions in this paper. We are particularly grateful for the loyal support, work and ideas of George Pasdirtz.

## Methodological and Theoretical Problems of Measurement in Preparing Historical Time Series Data for Periodization

To gain the fullest benefits of data reduction and explanation that periodized analysis can give, some attention must be paid to properly preparing the time series data. A number of problems occur in dealing with historical time series, and the manner of dealing with them will have a significant impact on the success of the later analysis. While not an exhaustive list, standardization, the number of time points and frequency of measurement, and the handling of missing data are among the most common difficulties.

### *Issues in Standardization*

Any attempt to periodize is, at least implicitly, an effort at comparative analysis either across time within one nation or across nations. For this reason it is necessary to standardize the time series so that the generalizations sought are less bounded by a particular space and time. Where it makes theoretical sense to do so it is also advisable to standardize in a way that reduces the series in question to percentage figures. Such series are easily interpreted and are relatively easy to manipulate. Standardizing the time series as percentages also has the advantage of yielding readily interpretable unstandardized beta coefficients when regression models are used in analysis.

The use of standardized data in addition to facilitating comparative analysis provides a control that protects against false periodization. For example, the period following the First World War in France would be incorrectly classified as one of decline in education if the sheer number of students were considered. Expressing students as a percentage of the population 6–19 years of age, however, leads to an opposite conclusion because the decline in the number of students was less proportionally than the decline in the actual number of young persons.

In any comparative analysis it is necessary to choose a theoretically important variable on which to standardize the series of interest. Governmental expenditure for education can be expressed as a percentage of G.N.P. or as a percentage of all government expenditure. We have applied several periodization techniques to time series on government expenditure for education and for health, welfare, and social security as a percentage of G.N.P., which is particularly appropriate to theories about societal resource allocation.

The choice of a denominator should be tied to the theoretical questions addressed, but within this structure the choices made can affect the generalizability of the analytic results. For example, Labour Party votes as a proportion of all votes cast in

a meaningful standardization for Britain and a few other countries at a few points in time. If the theoretical question to be addressed is a larger one a more generalizable choice is desirable. For example, if the theory is about the relationship between working-class political participation and resource allocation, a wider range of parties including all those „left of center“ is necessary.

### *Number of Time Points and Frequency of Measurement*

For any technique of periodization the length of the series and the frequency of measurement need to be taken into account. The series must be long enough to allow for significant change to have occurred and measurement must be frequent enough so that the timing of the changes can be captured accurately.

The questions of length and frequency are linked to the nature of the theory and the nature of the phenomenon under study. In the study of resource allocation and modernization in Western Europe we have chosen to take yearly measures for at least one century. For some phenomena such as urbanization, political development, and income distribution less frequent measurement is necessary. However, there are a number of important variables that require more frequent measurement, for example, strikes, balance of trade movements, and immigration. To correlate a rapidly changing series, such as strikes, with a slowly changing one, such as urbanization, requires that urbanization be measured as frequently as strikes.

### *Handling Missing Data*

The larger the number of cases, the longer the time span covered, and the more frequent the measurement the more likely one is to have difficulties with missing data. Conventionally, missing data is estimated by linear interpolation and extrapolation. If correlational and linear regression techniques are used for periodization, however, linear estimates of missing data will bias the results. The use of non-linear estimation techniques will avoid these biases.

We prefer to use a cubic polynomial spline method to estimate missing data. This method fits a non-linear function to each set of three „real“ observations and generates estimates for the missing data between them. For example, the proportion of the labor force in agriculture in France is directly observed only at the census that are, in most cases, five years apart. The polynomial spline used the three observations that occur over a fifteen year period to fill in the missing years. If the

trend in the real data over the period is linear, the estimated data will also be linear; if the trend is non-linear, the polynomial spline will produce a smooth, but non-linear, series.

We might point out as well that even qualitative estimates are better than nothing at all when attempting to estimate missing periods. This is especially important because the more frequently one can at least estimate time points the better any interpolation program will work. If one connects a series of qualitative estimations by non-linear interpolations is one constructing a house of cards? The answer is, not really. Different estimations and interpolations can be checked for the general plausibility both from the standpoint of the different periodizations they produce as well as their pattern of association with other variables. This is a variation on the theme of counter-factual history. If the estimated time series works reasonably well in data analysis and the same data analysis has been demonstrated to work in other countries or from historical eras then we can assume we have not done great violence to historical reality.

For example, union membership in Italy during the past World War II period has been a secret. We have used some „soft“ estimates based on attendance of delegates at conferences and interpolations between these estimates. English public education statistics in the 19th century are very misleading without some attempt to estimate the private sector. If one examines literacy rates in Britain, and especially England, they are as high or higher than those in France suggesting that not all students are accounted for in the official records. Qualitative reports indicate there were a large number of small, private, and usually for profit schools especially in the urban centers. Not to estimate them, however crudely, will mislead any technique of periodization.

In summary our rules for preparing data for periodization require the following procedures: (1) standardize and employ percentages whenever possible; (2) generalize the variables as much as possible, even if the analytical interest is relatively „narrow“; (3) obtain at least yearly measures for a *longue durée* of a century or more; (4) use non-linear estimation procedures for missing data; (5) attempt to use qualitative sources and reasonable guesses to supplement official records.

## Periodizing with a Single Variable

The coherence and explanatory power of periodized analysis can be achieved by examining even movements in a single time series. Economic historians, in particular, have effectively utilized this approach to identifying periods that make intelligible such diverse problems as long-run economic growth and business cycles. The identification of periods by either changes in the rate of change or by deviations from a

function of time are two such techniques that can be readily extended to sociological and political variables. Each of these two procedures has particular strengths and particular weaknesses, as are illustrated below.

### *Changes in the Rate of Change as Periods*

One of the basic questions in examining a single time series is the identification of periods within time-point to time-point changes (slope of the variable plotted against time) which are similar and different from other periods. That is, one seeks to decompose the series into periods not by the level of the process, but rather by movements in the trend. As a hypothetical example, we would *not* make the statement „real G.N.P. per capital was higher in the 15th century than in the second half of the 16th“, but rather „both the 15th and late 16th centuries were characterized by stability in real G.N.P. per capita, while the first half of the 16th century is a period of decline“. Statements of the latter type lead us closer to an examination of processes — dynamics and statics — and away from pure description.

To identify periods by the techniques of changes and changes in the rate of change, one simply takes the time series in question and expresses it in time-point change scores ( $t_2 - t_1$ ) or in percentage rate of change scores ( $(t_2 - t_1)/t_1$ ). In the case of relatively homogeneous time series, as in the examples below, transformation of the data are often unnecessary. Simple examination of the series by eye to identify trends and turning points will suffice. What is important is a sensitivity to the slope of the series (change), and changes in the rate of change (inflection points where the slope becomes different).

The demand and supply of public education expenditure in Britain, France, Italy, and Germany over the century 1870 to 1965 provides a convenient example of the technique. If one believes that an important cause of resource allocation to education is the demand for it, as is the assumption in economics, it is very important to be able to measure and periodize this variable.

One way to measure demand for public education expenditure is to use the number of students and the cost of educating them. As the proportion of persons ages 6–19 enrolled increases, and as the cost per student increase, the need for public expenditure grows. The cost of primary, secondary and higher education differs because of varying teacher-student ratios, the size of plant, and technological intensity. Clearly educating 100 college students is not the same as educating 100 primary students. The cost of secondary and especially technical and vocational school students is much higher than the cost of educating primary students. Thus an arithmetic increase in proportion of the cohort being educated often represents results in a geometric increase in demand for government expenditures.

Our solution to this problem as suggested to us by George Pasdirth is to multiply the per student cost of education at the previous time-point by the number of stu-

dents in the current year, and then divide this sum by the G.N.P. The resulting figure is a demand for a certain percentage of the gross revenue of a society to be allocated to education:

$$\text{demand} = \frac{\text{number of students} \cdot \text{cost per student}}{\text{G.N.P.}}$$

There are, of course, several things wrong with this measure. While it is sensitive to cost increases and sudden population increases, it is less sensitive to the increases in particular school populations. This can be measured if the total school population is disaggregated into separate pools each of which has different per capita cost. In no way do the measures really distinguish between a high quality and a low quality educational system, but this is not its intent. The objective here is to measure demand or need as a percentage of the total resource pool ( G.N.P. ).

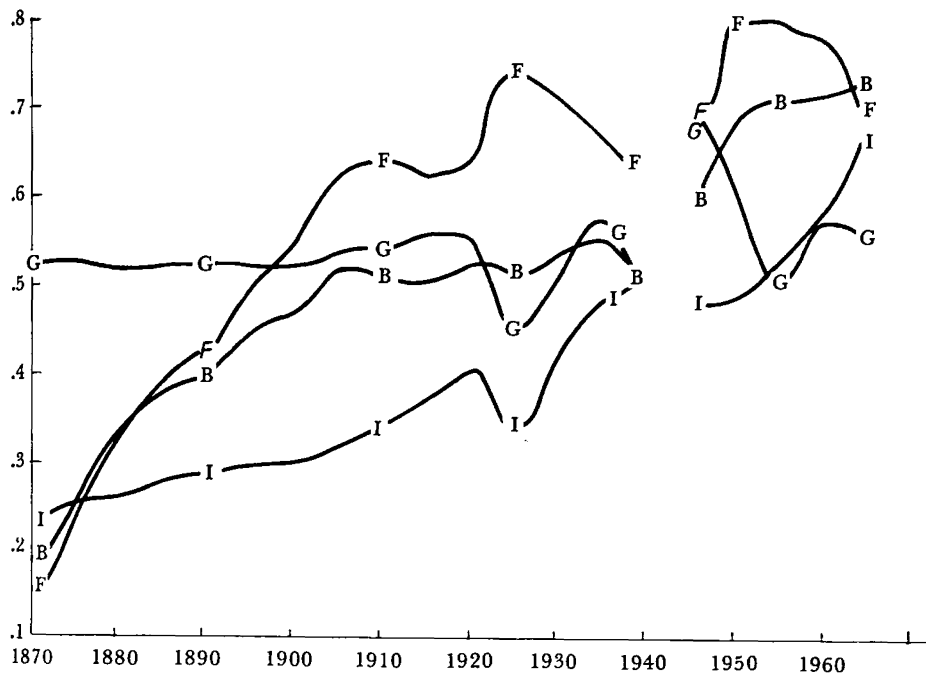
The more serious objection is that by taking only students in schools, we ignore children who would like to be in school but are not, either because of lack of facilities, inability to pay the fees, or the demands of rural work and the like. We are aware of this deficiency and are exploring several idealized models of growth curves in demand that could then be employed to estimate „true demand“.

Demand for public education expenditure for Britain, France, Italy, and Germany (territory of the Bundesrepublik since 1946) is plotted in Figure 1. These trace lines show some distinct pattern. German demand was remarkably stable with only minor fluctuations associated with the economic crises of the early 1920s and the post World War II bulge. Essentially our end point (1965) is little different from the 1871 starting point. France began at a considerably lower level than Germany and displayed moderate growth up to 1925, interrupted briefly by World War I. During the period from 1925 through the Depression and up to World War II demand declined. In the post World War II era demand expanded sharply to 1950 then declined to 1965. Britain started low, though figures include only the public sector and would be somewhat higher if the private sector was included, but had rapid growth up to 1905. From 1905 to World War II there was essential stability despite the establishment and slow growth of public secondary education. Taking the post-world war growth into account, the British pattern is quite like that of France since the turn of the century. Italy started at the same place as Britain, had slower but exponential growth until World War II, and rapid post war growth.

The kind and number of periods we observe are the following: there is one for Germany and it is a time of long-term stability; demand trends identify three periods for France: 40 years of moderate growth from 1870, followed by three decades of stability, and then a period of acceleration and decline after the Second World War. Great Britain experienced rapid growth from 1870 to 1905 followed by stability to the eve of the Second World War and, as in France, rapid acceleration and stagnation in the post-war years. In contrast to the other nations, Italy has known a moderate growth for fifty years from 1870 to 1920, and thereafter a very rapid increase up to 1965.

This periodization identified by the trace lines naturally raises the questions of

Figure 1: Demand for Education



explaining the differences among periods in the rate of change and the timing of changes in the rate of change that constitute the critical demand periods in the four societies. In the trace lines of Figure 1, zero slopes may be interpreted as stability, non-zero slopes as change. More important, however, are the places where the slopes themselves change. It is these inflection points that require the most emphasis in explanation: what started the periods of growth in the first place, what halted them, and how similar are the trajectories of the four countries?

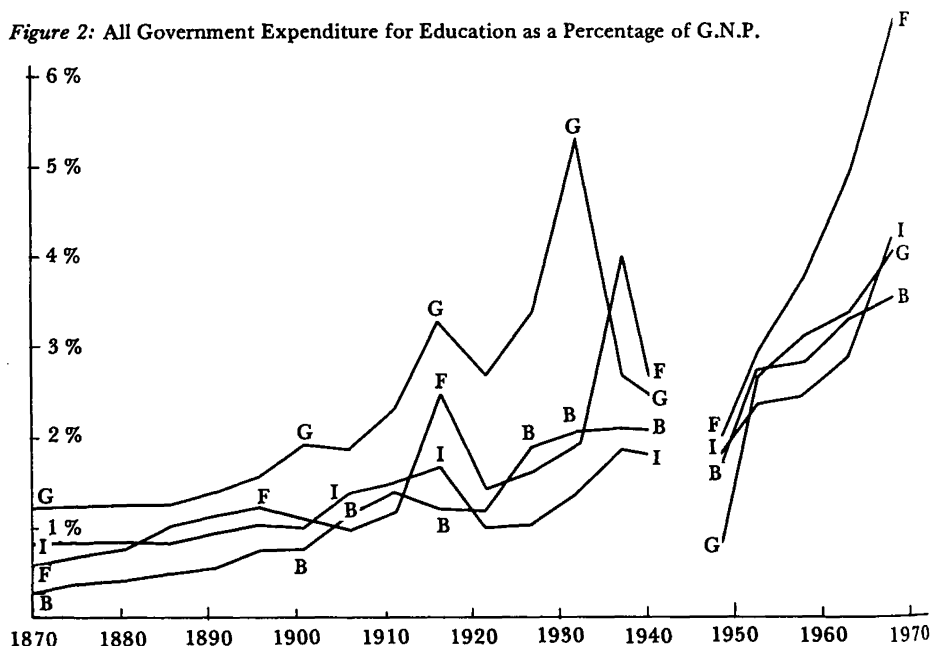
Substantively, the timing of changes in the rate of change is most interesting. Historians such as Antoine Prost have accounted major attention to the legislation which in a nation structures the goals and commitments of the society. Such specific events, however, do not account for all the subsequent movements in educational demand. Similarly, although much attention has been given to post Second World War acceleration in demand for education, upturns in demand for education occur in Germany and Italy before the war. To periodize by legislation or war is to miss some of the inflection or take-off points.



Descriptively, our procedures for periodizing a single series have raised historical questions that run counter to some of the perceptions of experts in the area. But note that these results would not be obtained without the use of time series for a *longue durée*, standardized in a consistent way, and juxtaposed vis-a-vis other countries in a somewhat similar state of modernization and of economic development.

Perhaps another example is necessary. In Figure 2 we plotted supply of education expenditure, that is the percentage of G.N.P. spent in education across all government levels in the four countries. For Britain, there were essentially two periods. Very slow growth up to World War II and after the war steady and more rapid growth. France has three distinct periods. Slow and undulating growth until a take-off in the 1930s, a violent swing associated with the war, and another take-off in the 1950s. Italy had a pattern somewhat like that of Britain. There was slow growth with some undulations in the interwar period, and then acceleration in the post World War II period. Germany again had a distinctively different pattern. There is exponential growth from 1870 to 1930 with only minor reversals, a sharp decline in the Nazi period, and rapid growth since the Second World War.

Figure 2: All Government Expenditure for Education as a Percentage of G.N.P.



It is worth noting that there is not a simple overlap between the periodization in demand (Figure 1) and supply (Figure 2). There is a good fit for Italy and Great Britain, the periods identified in demand and expenditure closely correspond. In Germany stable demand does not correspond to increasing expenditure. In France rapid post Second World War expansion of expenditure is not complementary to stable or declining demand.

These results encourage some observations on the periods that are delineated. First, we observe that the meaning of war and of depression may be quite limited when viewed in the perspective of a hundred years. Many studies of the economic growth rates in various countries indicates that they have a characteristic rate of growth. Depressions may interrupt this. But once the next boom is over, the long term average remains about the same. Here we see essentially the same phenomenon: a tendency for government expenditures on education as a *proportion of G.N.P.* to grow at a constant rate irrespective of the major discontinuities in the biographies of these four countries.

Second, we have created a new kind of datum. In so far as each country has a characteristic growth rate, then this becomes another analytical problem. From a comparative perspective what is interesting is the close correspondence between France, Britain and Italy for almost the entire time period. It is only in 1960 that France becomes quite different from the other two countries. Especially when one juxtaposes these findings vis-à-vis the different growth patterns in demand, one is struck by the similarity of government responsiveness in these three countries.

Third, if one examines the yearly changes more closely, another new datum is created: the characteristic way in which the changes in slope occur. In some countries as noted above, the slope changes are sharp, fitting more a step function pattern. In other countries, the slope changes are less distinct.

We have indicated how periodization can be accomplished by remembering that periods of constant slope are periods of continuity, and that the changes in slope represent discontinuities. We have observed that the starting points of new periods do not necessarily correspond with the dramatic events by which we frequently mark past time. This means that the root causes may best be found elsewhere.

### *Regression on Time Functions as Method Identifying Periods in a Single Time Series*

One of the major difficulties with using changes and changes in the rate of change to identify periods is the sensitivity of method to short-term fluctuations. If the series under examination does not display relatively smooth patterns, as do educational demand and supply, it is often very difficult to decide when significant „changes in the rate of change“ have occurred.

Econometricians have developed a number of techniques for dealing with this

problem, including smoothing to minimize short-run fluctuation, and seasonal adjustments to remove cyclical short-run movements. The relatively sophisticated techniques of spectral analysis and Box-Jenkins and Box-Tiao are designed specifically to identify and/or remove such „noise“ from the basic trends of a series.

A simpler approach is to regress the time series in question on a function of time and examine the residuals of the regression. When this is done, the pattern of residuals may be used to identify pure „noise“, cyclical fluctuations, and homogeneous sub-periods around the long-run trend. To use this technique implies a focus on two major questions. First, we must answer why a particular function of time has been used to define the major trend in a series (linear, logarithmic, sine wave, logistic, etc.). Secondly, we must focus on the analysis of periods of homogeneous residuals in terms of what may cause the periods to deviate from the trend. This method then leads us into the questions of why there is a basic trend, and helps us to identify the forces that cause deviations from the trend.

An illustration of this form of analysis is in order. Government expenditure for education as a percentage of the G.N.P. is again helpful. Examining the fit between a simple exponential function of time and the actual trend in education expenditures illustrates the way in which this technique may be used for identifying periods relevant to questions of deviation from long run trends.

We begin by supposing that the basic time function of the government education expenditure share in G.N.P. is a logistic (S-shaped) curve, but that in the time period of interest only the early exponential portions are observed. In order for deviations from the exponential pattern to be meaningful, a theoretical justification of the function is necessary. For the purposes of this illustration, let us suppose that education expenditure as a proportion of the national product grows logistically because parents who obtain education desire education for their children; the process reaches an upper bound when all children receive the maximum amount of education consistent with the maintenance of other societal functions.

In Figure 3 the trends of actual education of time for Britain and France are shown. Figure 3 raises a number of interesting questions as well as identifying periods developed through residual analysis within each series that are worthy of more detailed analysis. We are struck immediately with the problems of why the value of the French exponential parameter is greater than the British; we are also struck by the relatively small deviations around the trend in Britain and the relatively large swings in France.

In terms of periodization, deviations from the exponential time trend identify 1900–1910 and 1920–1940 as periods of more rapid expansion than expected in Britain. The periods 1910–1920 and 1946–50 require an explanation of the retardation that escapes notice when analysis is limited to changes in the rate of change. In France the periods 1895–1910, 1920–1930, and 1946–1950 display lower than expected levels while 1910–1920, 1930–38, and 1950–1965 exceed the expected values. In terms of period analysis, this examination of residuals from a time trend indicates the times where we are most likely to find evidence of forces at work that act to retard or enhance basic structural dynamics.

Figure 3: All Government Expenditure for Education as a Percentage of G.N.P. —  
Britain and France Actual and as an Exponential Function of Time



In summary, the analysis of a single time series can be used to identify periods that have substantive meaning. The method of periodization used is closely tied to the type of question that one wishes to investigate. Using changes (slopes) and changes in the rate of change (inflection points or second derivatives) is most helpful in identifying the causes of stability and change. Analysis of the residuals from a regression of a variable on a function of time is most helpful in identifying factors that modify the effects of an underlying dynamic process. Differences among nations or across time in the parameters of the underlying time function provide important clues to institutional differences.

The use of either approach moves one away from the temptation to perceive dramatic events as bench marks. The techniques do more than this, however, in that they raise a whole series of questions that require systematic explanations in terms of general theory. These methods allow one to note similarities and dissimilarities among countries and eras and thus speak to the basic concerns of both social science and comparative history.

## Periodizing with Multiple Time Series

Most historians and sociologists perceive periods not so much as distinctive movements in one variable, but rather as sets of time points within which the relationships among many variables are homogeneous and different from their interrelationships in other sets of time points. If one takes a large number of time series, describes each of them by periods, and seeks to identify correspondences, the sheer amount of descriptive material is overwhelming. If the relationships among the variables, that is their parameters, remain constant over time, there is no difficulty. In this case a single set of structured equations adequately describes the entire body of data. If relationships among variables change over time, however, a single set of equations is inadequate and a new method of periodizing must be found.

Some very sophisticated techniques exist, based on variations of factor analysis, cluster analysis, spectral analysis, Box-Jenkins and Box-Tiao techniques. Again, however, we advocate a simpler approach more in keeping with the current state of most sociological-historical theory and data.

### *The Use of Parameter Changes in the Delineation of Periods*

In the cliometric work of Williamson, the parameters estimated for the 1870s for a general equilibrium model of economic growth of the United States work reasonably until World War I<sup>1</sup>. If one examines the large literature on the status attainment model one is struck by the changes in betas across time and samples. One could add other evidence, but all of it suggests that parameters — the relationships among multiple variables — change over time. André and Delorme for example, found that elasticities between changes in student populations and increases in government expenditures were different in different periods<sup>2</sup>. However, their designation of periods was conventional (1871–1914, 1921–1939, and 1946–1971). Rather than using events to identify periods, we would like a procedure that allows the periodization by multiple variables to emerge from the analysis of their interrelationships.

For example, a minimum identification of the variables at play in the growth of government expenditures for health, welfare, and social security would include real

<sup>1</sup> Williamson, Jeffrey G., *Late Nineteenth-century American Development. A general equilibrium history*, London 1974.

<sup>2</sup> André, C., and Delorme, R., *Etude analytique et numérique des tendances significatives et des facteurs explicatifs de l'évolution des dépenses et recettes publiques Françaises au cours de la période 1870–1970*, Paris 1973.

G.N.P. per capita, labor force structure, and political polarization as measured by the votes received by left and right parties. A number of other independent variables, of course, might be identified, but for illustrative purposes of periodizing with multiple variables, this list will suffice. How then can the relationships among these variables be periodized to aid in the explanation of governmental expenditure?

Our recommendation is that we regress government expenditures for health, welfare, and social security as a percentage of G.N.P. on indicators of the set of independent variables and examine the residuals. Periods are identified by clusters of residuals above or below the line that summarizes the „average“ relationship among the variables over the entire time period<sup>3</sup>.

In Table 1, the summary statistics are presented of the regression of government expenditure for health, welfare, and social security as a percentage of the G.N.P. on indexes of three sets of independent variables for Britain, France, Germany, and Italy over the period 1875–1965. The „resources“ index includes real G.N.P. per capita and labor force structure; the „right“ index includes the percentage of votes for conservative and centrist political parties, the proportion of the labor force in agriculture, and the average size of firms; the „left“ index includes the proportion of the labor force in unions, and votes for Socialist and Communist parties.

*Table 1: The Partial Correlation of Resources, Right Power Base, and Left Power Base on Government Expenditures for Health, Welfare, and Social Security as a Percentage of G. N. P.*

	Resources	Right	Left	R	R <sup>2</sup>	DW*
Britain	–.27	–.66	+.57	.98	.96	yes
France	+ .41	–.30	+.22	.92	.85	yes
Germany	+ .10	–.39	+.32	.87	.75	yes
Italy	+ .74	–.10	+.41	.92	.84	no

\*Significant negative autocorrelation as measured by the Durbin-Watson statistic.

As can be seen from Table 1, the model works quite well in each of the four countries, accounting for between three-fourth and ninety-five percent of the variance ( $R^2$ ). Admittedly, the correlations are somewhat inflated because of significant and negative autocorrelations in three of the four nations. In Britain and Germany the political variables are more important than the economic, in France and

<sup>3</sup> Details of definition and measurement for this specific example are described in Hage, Jerald, and Hanneman, Robert, *The Growth of the Welfare State in Four Western European Societies, 1870–1965*. Institute for Research on Poverty Discussion Paper, University of Wisconsin, Madison (forthcoming).

Italy the reverse is true. This finding is interesting because in most cross-sectional studies political variables appear to be of lesser significance.

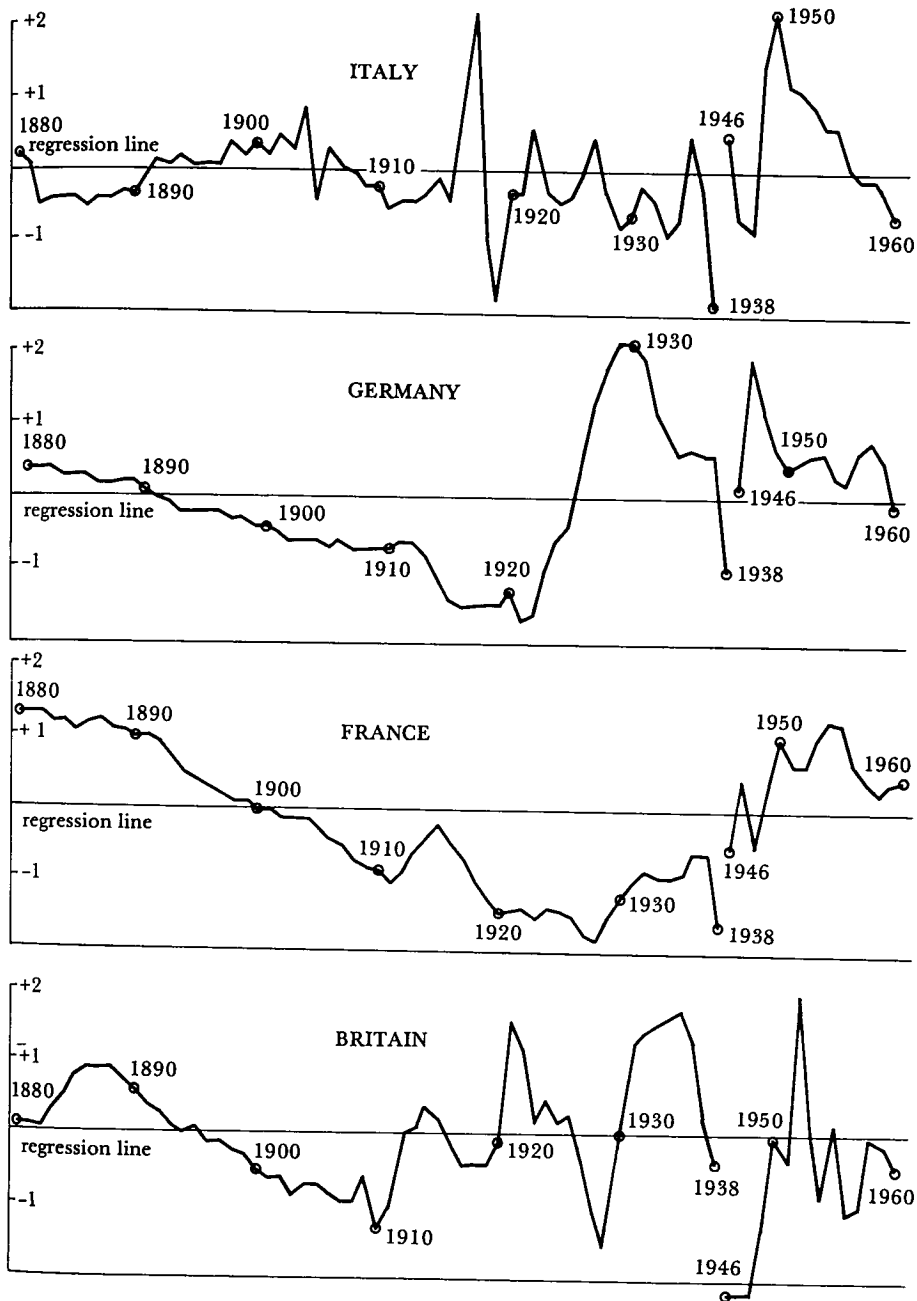
With these very high correlations, one might assume that a quite adequate job had been done in explaining the extent of government expenditures. However, this ignores the fact that for some periods the model might work very well and in other periods the model may break down. When and where the model does and does not fit well is an important datum, and can be seen by examining the residuals — that is, the goodness of fit between the predicted government expenditure in a year and the actual expenditure. Residuals of the model for each nation are presented in figure four. In Figure 4, a positive residual indicates that the actual expenditure in the year is greater than the prediction from the model, a negative residual that the expenditure is less than the prediction.

In Italy, from a regression point of view, the results have a nearly ideal pattern. Residuals do not cluster above or below the line for sets of sequential years, and consequently there is little autocorrelation. Most important, there are not long time periods when the residuals are either positive or negative. With the exception of the 1950s and 1960s, where there are some patterns among the residuals, there are, by this method, no distinct periods for Italy, and the parameters of the model as originally estimated are a good characterization of the relationships among the multiple variables for the entire century.

In Germany, the exact opposite is the case. The model accounts for less variation in Germany and there is a significant negative autocorrelation. Here we find two long periods and two short ones that correspond not only to the history of social welfare effort in Germany, but also, to some extent to political periods. The first period, 1878–1891, corresponds closely to the time of the Bismarckian welfare and social security legislation. From 1891, one year after Bismarck was dismissed, a long period continues through the First World War and the hyperinflation during which expenditures are less than the predictions of the model. This would appear to define a distinct historical era. From the mid-twenties until the beginning of the sixties, one finds another distinct period, one that includes Weimar, the Third Reich, and the Adenauer post-Second World War years. Throughout, more was expended than is predicted by the model. Finally, there is another short period that starts with 1960 and continues to our last data point at 1965.

In France there are three distinct periods, each of very long duration, and each raising interesting historical questions. Although the amount of variation accounted for is the same as Italy, the patterns of residuals indicate the existence of distinct periods. The first period goes from 1878 to 1900, during which expenditure is greater than predicted. From 1900 until about 1950, the reverse is the case, with normal fluctuations about the two wars. The third period starts at about 1950 and continues to the present.

Figure 4: Residuals from Regression





The case of Britain is in certain respects more like that of Italy than like the other two nations. In this country the model accounts for more variance than in any other of the four societies. There are more swings in the residual pattern and less distinct demarcations of historical periods, but some exist nevertheless. There is one long period from 1878 to 1896 and then another up to 1911. The interwar period is one when there is in general more expenditure than predicated. What is perhaps most surprising is the long period after the Second World War when there is less expenditure than the model predicts. Since this is the era of a number of Labour governments and „the growth of the welfare state“, this seems surprising. The results suggest, however, that the welfare state did not expand as much as would have been the case if the conditions of resources and right and left strength had the same impact in the post-war period as they did over the entire period. Finally, there is a short period in the 1960s in Britain when expenditures again exceed the expected.

The examination of the residuals from a single regression model (and more complex models could be used in the same fashion) enables us to identify periods where the model does not fit. Due to limitations of space, the analysis of this example will not be pursued further. It should be pointed out that, once the initial set of residuals have been examined, the model may be re-estimated for shorter periods, so that the patterns among the residuals may be summarized by sets of parameter estimates. Once this is done, the logical questions become: why are the parameters different in different periods?, and what are the causes of the parameter shifts? This method not only allows us to identify periods of time when, for example, resource constraints are more important than political process in the explanation of social welfare effort, but also forces us to explain why this is not always the case. We are obliged to try to understand the conditions under which one „model“ of the behavior in question works better than another, and how a social system may move from one set of dynamics to another.

## Conclusion

We have tried to remain faithful to both the topic of how to create meaningful historical periods and finding new sources of data. The problem is usefully seen as both a descriptive and analytical one, both theoretical and methodological. Periods may be identified by changes in slopes or deviations from a time trend in a single variable, or they can be determined by changes in the parameters defining the relationship between two or more variables.

A consequence of the periodizations established here is that, in some instances, traditional chronology is sustained, and in others it is inadequate. Critical attention is drawn to the times when societal performance shifts. The multiple variable periodi-

zation technique highlights breaks in the consistent interplay of variables and accents the critical changing weights of variables. The much celebrated autonomy of politics is reinforced at times with regard to public expenditure, and at other times it must give place to other social phenomena such as demand and resource availability. This experience enhances the significance of the periods in public expenditure for Britain, France, Italy, and Germany where simple models do not fit and calls for historically specific evaluation. Periodized analyses of this sort indentifies the actual periods presumed to occupy historians and likewise the complexities best approached by social science analysis. The concerns of historians and social scientists become then complementary, and the generalizations and theory they seek are not competitive but mutually supportive.

## Analysis of Change in Discrete Variables\*

### 1. Introduction

Longitudinal data are analyzed using a variety of techniques and methods in the various social and behavioral sciences. Over-time data comes in many forms — as panel data, time series, and event-histories<sup>1</sup>. Different disciplines have tended to focus on one particular type of over-time data. Econometricians have concentrated on time series, demographers on a particular form of event-histories, sociologists on panel data, psychometricians on change scores. Further, the different disciplines have specialized in particular methodological problems — econometricians in problems of estimation, especially in problems of time dependent errors, psychometricians in the reliability of change scores, and, in classical panel analysis, sociologists have concentrated on developing measures of causal influence. The result is that longitudinal methodology is a confusing affair. Some problems have solutions, others equally important do not, and it is often difficult to see the relevance of a technique for a problem if this technique has been developed in another discipline with a different research tradition.

There is then a need both for codification and for remedying some of the uneven development of existing longitudinal methodology. One set of problems for which longitudinal methodology seems in particular need of attention is composed of those encountered when analyzing change in discrete or categorical variables.

\* The research reported here was supported by funds granted to the Institute for Research on Poverty at the University of Wisconsin by the Department of Health, Education and Welfare pursuant to the provisions of the Economic Opportunity Act of 1964.

<sup>1</sup> Event-history data are longitudinal data where the exact timing of events is known. They are thus continuous time records of events like job shifts, residence shifts, etc., when the units of analysis are individuals. Life-history data are event-history data. For methodological purposes, the important feature of life-history data is the information on the timing of events, not the coverage of people's lives. Further, in some instances, information on the timing of events may be obtained with other designs than the life-history study design. Hence the term event-history is preferred.

Although such variables are often employed by the softer social sciences, there does not exist a readily available set of techniques and methods for the analysis of change.

Until recently there was in fact very little that could be done with categorical variables other than computing percentages. This situation has changed dramatically in recent years. Powerful techniques for log-linear and multiple classification analyses have increasingly become available<sup>2</sup>. These techniques may of course also be applied to change data, treating time as any other independent variable, and analyzing the data using the approach applied to cross-sectional data. These techniques of course also may be applied to over-time data in the same manner as they are applied to cross-sectional data. The over-time variation is then treated in the same manner as the cross-sectional variation among individuals (or other units of analysis).

Using log-linear and multiple classification analysis with over-time data on categorical variables corresponds to the use of regression techniques with algebraic (usually linear) models with cross-sectional and over-time data on continuous variables. In both instances the interest is in estimating the relationship among variables and in both cases the over-time variation is treated in the same manner as the cross-sectional variation in variables. Much social science research has the estimation and interpretation of relationships as the primary objective. These *ad hoc* techniques are then appropriate and the methodology, despite very different estimation techniques, is conceptually similar whether log-linear or multiple regression techniques are applied.

The application of *ad hoc* statistical techniques makes assumptions about the form of the relationship among variables, usually that they are linearly related. Such assumptions are rarely tested and even more rarely justified in terms of substantive consideration of the process under study. Nevertheless, the assumption made may be empirically and conceptually inadequate. This may lead to misleading inferences and limit our ability to understand fully the processes being analyzed. On cross-sectional data there is, however, very little that can be done since the unfolding of the processes that generate observed relationships among variables cannot be observed. On over-time data it is possible to study directly the change processes that generate observed outcomes. However, when over-time variation is treated as cross-sectional variation, this opportunity for obtaining a better understanding of how observed outcomes are generated is missed. Direct study of change is needed. This paper will advocate such an approach to the study of change in discrete variables.

To study change it is necessary to identify the components of change. The first part of this paper will identify these components. The second part of the paper will then briefly outline some strategies for the causal analysis of these components.

<sup>2</sup> A comprehensive survey is provided by Bishop, Y. M. M., et al., *Discrete Multivariate Analysis: Theory and Practice*, Cambridge 1975.

## 2. Conceptualizing Change

The concern in longitudinal methodology is with description and analysis of variables that are functions of time. To identify the tasks involved it is necessary to have a representation of the change process that identifies the quantities that should be estimated in empirical analysis. In other words, a conceptualization of the change process should be given in a mathematical representation. The classic approach to the mathematical analysis of change is the one represented by calculus. It applies to variables that are continuous, i. e. variables that can be represented by real numbers. Though the concern in this paper is with discrete variables, the continuous variable treatment serves as a model, and will be briefly outlined.

It seems natural to represent change as the difference in values of the variable of interest obtained over some time interval. Denote the time dependent variable  $y(t)$ . The difference  $y(t_2) - y(t_1)$  observed over the interval  $t_2 - t_1$  would be the quantity of interest. Presumably this difference is brought about by some causal variables, possibly including time, that act on  $y(t)$  in a certain way. In descriptive analysis the objective is to specify the resulting time variation in  $y(t)$ . In causal analysis we go further and attempt to specify the various causal forces acting on  $y(t)$ , and estimate their influence. In other words, for causal analysis it is necessary to specify: (1) the mechanisms that bring about change, and (2) to assess the causal influences transmitted by these mechanisms.

### 2.1. *Specifying the Mechanism of Change*

The specification of the change mechanisms depends first on the timing of change. If  $y(t)$  changes continuously in time so that  $y(t)$  is continuously differentiable with respect to time over the interval of interest, relating  $y(t_2) - y(t_1)$  to  $t_2 - t_1$  presents the problem that as  $y(t)$  changes, so does  $t$ . The classic solution is to focus on the change in  $y(t)$  obtained in an infinitesimal small interval of time<sup>3</sup>. This conceptual abstraction makes it possible to relate change to the value of time (and other variables) rather than to intervals of time. Hence we focus on the quantity  $dy(t)/dt$ , i. e. the instantaneous rate of change in  $y(t)$ . The specification of the dependency of  $y(t)$  on time and other variables may then be carried out in a differential equation:

$$\frac{dy(t)}{dt} = f(\underline{x}(t), \alpha, t) \quad (1)$$

<sup>3</sup> See for further discussion Coleman, J. S., *The Mathematical Study of Change*, in: Blalock, Hubert M., and Blalock, Ann B. (eds.), *Methodology in Social Research*, New York 1968.

where the vector  $\underline{x}(t)$  represents causal variables, possibly including time and  $y(t)$  itself, and the vector  $\alpha$  represents a set of parameters.

The specification of  $f$  in the differential equation should represent assumptions about how change is produced. Some simple examples will illustrate the strategy.

The simplest process is obtained assuming that  $y(t)$  changes by a constant amount in each small interval of time, or:

$$\frac{dy(t)}{dt} = k \quad (2)$$

A slightly more complicated expression, that is a useful representation of many processes, assumes that change in  $y(t)$  is dependent on  $y(t)$ :

$$\frac{dy(t)}{dt} = k + by(t) \quad (3)$$

If the quantity  $b$  represents a feedback either positive or negative. In many growth processes this feedback will be negative, and (3) describes a process where  $y(t)$  changes rapidly in the start of the process, but decreases as  $y(t)$  increases and eventually reaches zero at the equilibrium level of  $y(t)$ , where  $dy(t)/dt$  is zero. Though stable processes will have this property, there may be considerable interest in processes with positive feedback where the variables of interest will take an explosive course. One example is arms races leading to wars that have been modeled by Richardson in a simultaneous differential equation model with basic properties like (3) though mathematically more complicated<sup>4</sup>.

Since  $dy(t)/dt$  is a conceptual abstraction, differential equations cannot be used directly with empirical data. In order to estimate parameters and test the models it is necessary to solve the equations using methods of integration. For example, the solution to equation (2) is:

$$y(t) = y(0) + kt \quad (4)$$

where  $y(0)$  is the value of  $y(t)$  obtained at the start of the process, at time 0. The solution to (3) is:

$$y(t) = \frac{k}{b} (e^{bt} - 1) + y(0) e^{bt} \quad (5)$$

Expressions such as (4) and (5) may be used with empirical observations on  $y(t)$  and  $y(0)$  either for a set of individuals (or whatever is the unit of analysis) or obtained through repeated observations on the same individual. These formulations are necessary to test the models and estimate parameters, but the conception of the change process is given by the differential equation, from which the parameters derive their interpretation.

It is important to note that the solution (5) to (3) only holds if the parameters  $k$  and  $b$  are assumed constant over time and identical for all individuals. Failure of

<sup>4</sup> Richardson, L. G., *Arms and Insecurity*, Pittsburg 1960.

these assumptions of stationarity and homogeneity will result in models that do not describe the observed course of processes adequately. Failure of the assumption means that characteristics of individuals and/or time periods cause variation in the components of change. Such variation should be modeled. The specification of the sources of variation provide the desired information on the causes of change, as shown below.

The use of differential equations to mirror change processes depends on the continuous differentiability of  $y(t)$  with respect to time. If change does not take place continuously, but only after certain intervals of time, a different formulation is necessary. Change may then be modeled in a difference equation treating time as a discrete (integer) variable:

$$\Delta y = f(x(n), \alpha, n) \quad (6)$$

where  $n$  is used to represent time, often trials or other discretely occurring events. A different equation may be estimated directly, since the quantity  $\Delta y$  usually is observable. This is sometimes seen as an advantage, and difference equations are for example often used in economics because observations are obtained at fixed intervals of time (e. g. at yearly intervals). On the other hand difference equations will still need to be solved in order to study the over-time behavior of the process and test the models, and the standard methods of calculus are not available for this purpose. Further the conception of change, not the timing of observations, should govern the formulation of a model of change. This will usually dictate the continuous time formulation in a differential equation model.

## 2.2. Specifying the Causes of Change

The examples brought above were examples of models expressing the mechanism of change in time, but not the dependency on other variables. One useful way of introducing causal variables is to express the parameters of the models as functions of a set of independent variables. In equation (3), the quantity  $k$  may for example be written as a linear function of a set of exogenous variables, i. e.  $k = c_0 + c_1 x_1 + \dots + c_n x_n$ . This will result in:

$$\frac{dy(t)}{dt} = c_0 + by(t) + c_1 x_1 + c_2 x_2 \dots + c_n x_n \quad (7)$$

The solution to (7) is parallel to (5) with the linear expansion of  $k$ . It is important to note that if  $b > 0$  and as  $t \rightarrow \infty$  the solution to (7) will reduce to:

$$y(e) = -\frac{c_0}{b} - \frac{c_1}{b} x_1 \dots - \frac{c_n}{b} x_n \quad (8)$$

The equilibrium formulation of (7) is thus the simple linear model for a variable so often used on cross-sectional data. Note that the derivation from (7) shows that the

quantities —  $\frac{c_j}{b}$  that are the observed coefficients to the independent variables depend on the feedback term  $b$ . In other words, starting out with the model of change, equation (3) results in a formulation of the relationship among variables that may be observed in a cross-section in terms of the fundamental quantities that govern change. Only over-time data can identify these quantities, and only modeling change directly will specify the components of change. Over-time analysis that treats over-time variation as cross-sectional variation will not provide this information, as it will amount to using models such as (7) with time as an independent variable; an inappropriate conception if (3) governs the change process. (For further implications of this and other results of modelling change directly see Sørensen<sup>5</sup>.)

Writing parameters in simple change models as functions of causal variables is only a meaningful way of modeling the causes of change if it can be assumed that the independent variables are unaffected by  $y(t)$ , i. e. that there is no interdependence among  $y(t)$  and the  $x_i$  variables. If this cannot be assumed more complicated simultaneous differential equation models are needed to mirror the change process. These complications will not be discussed here.

The specification of the variation in quantity  $k$  of equation (3) in terms of the  $x_i$  variables also should make the model more empirically adequate since the heterogeneity in  $k$  is taken into account. Further modification may allow for time dependency, though the resulting models are quite complicated<sup>6</sup>.

Causal analysis of change processes then demands first a specification of the mechanism of change in a differential or difference equation. The causal variables may be introduced directly in the defining equation. In many situations it is, however, simpler to see the causal variables as acting on the parameters that govern change. This is the approach that will be suggested for the analysis of discrete variables, to be discussed in the remainder of the paper.

### 3. Conceptualizing Change in Discrete Variables

Analysis of change in continuous and in discrete variables differ in one all important respect. Change cannot meaningfully be represented as differences in the values of variables when the variable is discrete. Hence differential or difference equations cannot be used to represent the change process, and calculus cannot be applied directly to the variables.

<sup>5</sup> Sørensen, Aage B., Causal Analysis of Cross-Sectional and Overtime Data: With Special Reference to the Occupational Achievement Process, in: Weselovski, W. (ed.), *Social Mobility in a Comparative Perspective*, forthcoming.

<sup>6</sup> See Coleman, *Introduction to Mathematical Sociology*, New York 1964, for an example.



The problem is sometimes solved by treating discrete variables as though they have a stronger metric. It is, for example, common in sociology to treat the standard measures of occupational prestige as though they possess interval level metric, though they are ordinal measures. Similarly dichotomous variables are often treated in the same manner as continuous variables in regression analysis. This solution is, however, often conceptually unsatisfactory, and the obtained estimates have undesirable statistical properties. An alternative solution is to study change in discrete variables by „proxy“; by mapping the categories of the discrete variables onto a probability distribution. The probabilities provide the desired metric, and can be studied as change in probability distributions over the state space given by the categories of the discrete variable. Probability theory (of course often using a great deal of calculus) becomes the relevant mathematical language for the study of change, and the resulting models will be stochastic process models.

Change in continuous variables could also be studied by focussing on change in probability distributions and applying stochastic process models with continuous state space. However, the mathematical complications are considerable, and the mathematical problems often become serious with discrete state models. On the other hand, the use of stochastic process models is the only way of modeling change in discrete variables, and this, rather than a fundamental choice between a stochastic versus a deterministic conception of a process, seems to be the usual reason for the use of stochastic process models with discrete variables and deterministic models with continuous variables.

As with continuous variables, the timing of change determines whether the defining equation is a differential or a difference equation, and as describes above, these equations have to be solved in order to estimate parameters and test the models. However, solutions to stochastic process models, except the very simplest, are usually quite complicated and in fact often impossible to obtain (see for example the epidemiological models presented by Bailey<sup>7</sup>). On the other hand, because stochastic process models permit a microscopic analysis of the process of change, even very simple models may provide a wealth of information for the analysis of the various components of change.

Suppose now that the variable of interest is a dichotomous variable giving rise to a two-state system. Label the two 1 and 2 respectively. A unit of analysis, say an individual, is at a point in time,  $t$ , characterized by the probability  $p_1(t)$  of being in state 1, and  $p_2(t)$  of being in state 2, where  $p_1(t) = 1 - p_2(t)$ . The objective is to formulate the mechanism for change in  $p_1(t)$  and by implication,  $p_2(t)$ .

If change occurs continuously, a continuous time stochastic model is desired and should be defined in a differential equation model. Change in  $p_1(t)$  will reflect movement in the state space. Movement may either take place in one direction only — as when the two states refer to life and death — or, there may be movement in both directions — as when the states refer to a positive and a negative attitude. If movements in both direction take place, change will be governed by the probability

<sup>7</sup> Bailey, N. T. J., *The Mathematical Theory of Epidemics*, New York 1957.

of a move from state 1 to state 2 in an interval of time, and the probability of a move from 2 to 1 in the same interval of time. Denote  $q_{12} dt$  the probability of moving from 1 to 2 in  $dt$ , and  $q_{21} dt$  the probability of moving from 2 to 1. Assume further that these quantities are constant over time. Then the probability of an individual being in state 1 will change in  $dt$  according to:

$$\frac{dp_1(t)}{dt} = -q_{12}p_1(t) + q_{21}p_2(t) \quad (9)$$

where  $q_{12}p_1(t)$  is the rate of movement from 1 to 2 times the probability of being in state 1, and  $q_{21}p_2(t)$  similarly the rate of movement out of state 2 to state 1 times the probability of being in state 2. The expression is easily generalized to cover a larger number of states:

$$\frac{dp_i(t)}{dt} = - \sum_{i \neq j} q_{ij}p_i(t) + \sum_{i \neq j} q_{ji}p_j(t) \quad (10)$$

where the two parts of the right hand side govern respectively the outflow and the inflow from and to state  $i$ . For a  $k$  state system there will be  $k$  such equations. Assuming the  $q_{ij}$ 's constant, these equations can be solved to give the expression needed for empirical analysis. It becomes, in matrix notation:

$$\underline{P}(t) = \underline{P}(0)e^{Qt} \quad (11)$$

where  $\underline{P}(t)$  is the vector of probabilities at time  $t$ ,  $\underline{P}(0)$  the probability distribution at time 0, and  $e^{Qt}$  the matrix analog to  $e^a$  with  $Q$  a matrix of  $q_{ij}$ 's. This is the discrete state, continuous time Markov Model. Its application to social processes has been extensively discussed by Coleman<sup>8</sup>.

The discrete time analog to (10) is obtained from quantities  $r_{ij}$ , that are transition probabilities for moving from state  $i$  to state  $j$  on a trial. The typical equation for the change in the probability of being in state  $i$  on a trial will be:

$$\Delta p_i = - \sum_{i \neq j} r_{ij}p_i(n) + \sum_{i \neq j} r_{ji}p_j(n) \quad (12)$$

The solution to the set of different equations is, in matrix notation:

$$\underline{P}(n) = \underline{P}(0)R^n \quad (13)$$

analog to (11). Although discrete time processes are most often met in experimental situations, the discrete time Markov model is often applied to continuous time processes. Its advantage is mathematical simplicity. The distinction is often unimportant for prediction. However, for analysis the continuous time model often seems the most appropriate framework. One reason is that change can be further decomposed with continuous time models.

The quantities  $q_{ij}$  of the continuous time model give the rate of movement from state  $i$  to state  $j$ . It is often appropriate to conceive of this rate as resulting from the

<sup>8</sup> Coleman, J. S., Introduction.

occurrence of events randomly in time, and the outcome of events. Thus, in occupational mobility processes, a shift of occupation is the result of a job shift with a certain outcome, i. e. a possible shift of occupation. The occurrence of events and the outcome of events may be analyzed separately. Formally this means that the quantities  $q_{ij}$  may be decomposed as:

$$q_{ij} = \begin{cases} \lambda m_{ij} & i \neq j \\ \lambda(m_{ii} - 1) & i = j \end{cases} \quad (14)$$

where  $\lambda$  governs the occurrence of events, and the  $m_{ij}$ 's are the probabilities of moving from  $i$  to  $j$  given that event occurs.

With this decomposition, equation (11) can be written:

$$\underline{P}(t) = \underline{P}(0) e^{\lambda(M-I)t} \quad (15)$$

as the matrix  $Q$  of (11) =  $M-I$ , where  $I$  is the identity matrix. This formulation has been extensively discussed by Singer and Spilerman<sup>9</sup>.

In the simple continuous time Markov Chain, the occurrence of events is governed by a Poisson process. This means that the probability  $p_o(t)$  of no event occurring by time  $t$  will change according to the differential equation:

$$\frac{dp_o(t)}{dt} = -\lambda p_o(t) \quad (16)$$

The space for the Poisson process is a count of the number of events. The probability distribution corresponding to this state space is the Poisson distribution:

$$p_i(t) = e^{-\lambda t} \frac{(\lambda t)^i}{i!} \quad (17)$$

where  $p_i(t)$  is the probability that  $i$  event has occurred by time  $t$ . The mean of the distribution is  $\lambda t$ , a property that may be used to estimate  $\lambda$ .

Of considerable interest for analysis is the distribution. In a Poisson process this distribution will be exponential, with probability density:

$$f(s) = \lambda e^{-\lambda s} \quad (18)$$

where  $s$  stands for the time interval between events. The mean of  $s$  is  $1/\lambda$ , a property that again can be used in analysis of the occurrence of events.

The continuous time Markov Chain and the associated Poisson process for the occurrence of events are very simple. In fact the Poisson process is the analog to the simplest model for change in a continuous variable given as equation (2) with a constant increment in  $y(t)$  in each interval of time, and the Markov Chain is the analog to equation (3) where change is also assumed to depend on the current state of the system (in equation (3) on the value of  $y(t)$ ). These simple stochastic models may

<sup>9</sup> Singer, B., and Spilerman, S., Social Mobility Models for Heterogeneous Populations, in: Costner, H. L. (ed.), Sociological Methodology 1973-74, San Francisco 1974.

appear quite unrealistic models for change in discrete variables. They do, however, mirror the basic components of change in discrete variables, the distinctions between the occurrence of events and the outcome of events particularly important for analysis of change. Their appropriateness and one's willingness to live with their simplicity to some extent depends on the objective of the analysis of change, as the next section will describe.

#### 4. Objectives for the Analysis of Change

Models such as those described in the preceding section are introduced because of a desire to model the behavior of a process. This desire may reflect an interest in predicting the future course of a process, or an interest in formulating a theory of the process, or to provide a framework for a causal analysis of the components of change. Ultimately these three objectives may merge, but before the ultimate is achieved, different criteria for the usefulness of the models may be applied depending on which objective is emphasized.

If the objective is to predict, or if the objective is to formulate a theory, the primary emphasis is on the modelling task. The analysis of empirical data on change is carried out primarily to test the predictions from the model and validate its assumptions, not because of an interest in observed patterns of change and their empirical causes.

As a theory of a process the simple Markov model is quite uninteresting, and it has been repeatedly shown that the process does not predict well many social processes. The model's failure may have numerous causes, and an extensive literature exists on how to modify the simple model in order to improve its empirical or theoretical adequacy. Much of the literature on empirical adequacy addresses two problems: one is the problem of non-stationarity — that is, the fact that parameters change over time, the other is the problem of population heterogeneity — that is, parameters vary among individuals or whatever are the units of analysis to which the model is applied. Both non-stationarity and population heterogeneity will result in failure of the model to predict observed processes. Numerous solutions have been suggested in the literature that will improve the fit of the simple Markov model. They will not be reviewed here.

The discussion in the preceding section is intended to provide a point of departure for empirical analysis of the causes of change. Such analysis will focus on the sources of variation in the parameters that govern change, using continuous and discrete independent variables to account for this variation in the parameters that govern change in a manner analogous to the specification of equation (3) in equation (7). The utility of the simple models then lies in their identification of the

components of change. Non-stationarity and heterogeneity become of interest not because they are sources of failure of the models, but because they are the phenomena we would like to account for by causal variables. They are the objects of analysis, rather than something to get rid of.

## 5. Panel Versus Event-History Data

The representation of the Markov model presented in equation (15) suggests that analysis of change in discrete variables may focus on the variation in what governs the occurrence of events, and on variation in the  $m_{ij}$ 's that govern the outcome of events. However, separate analysis of the two components of change is only possible if the data provide the necessary information. Most data on change in discrete variables in sociology are obtained from panels. Panels are usually only observations at two or three points in time on a group of respondents. Such data can be used to estimate transition probabilities and from these transition rates may be computed.

However, since only a few observations are made on the process, information on the components of change will be very fragmentary. The resulting difficulties have recently been extensively analyzed by Singer and Spilerman<sup>10</sup>. With a large sample some analysis may be performed of variation in transition rates among sub-groups, but individual level analysis is impossible.

Event-history data are still rare, but far superior to panel data for causal analysis of change. With continuous observations on a group of respondents, waiting times between events may be directly observed in order to study variation in . Counts of the outcome of events may be used to obtain information on the  $m_{ij}$ 's. Event-history data thus provide much richer possibilities for analysis than do panel data, particularly for analysis of the rate at which events occur. The suggestions that follow for such analysis assumes that life-history data are used.

<sup>10</sup> Singer, Social Mobility; Singer, B., and Spilerman, S., Representation of Social Process by Markov Models, in: American Journal of Sociology, 82 (1976), pp. 1-54.

## 6. Analysis of the Occurrence of Events

Event-histories of the kind I am assuming will provide information on the timing of certain events and their outcomes. The histories may pertain to individuals and the events may be acts carried out by them such as a change of job or of residence. Or, the event-histories may pertain to societies, and events may be wars or elections (if elections can occur in any time interval). The purpose of the analysis would be to study the causes of variation in the occurrence of events.

With the Poisson process as the framework there are two ways of carrying out such analysis. One is to rely on the Poisson distribution and use counts of events to estimate the rate at which they occur, the other is to rely on information on waiting times between events.

If counts of events are relied on, the rationale is that the probability distributions over the state space given by the count has a mean that is  $\lambda t$ . Since  $t$  is known, a count of the number of events that have occurred to a person or a group of persons will provide the desired estimate. More precisely, we may, for example, carry out a count for each respondent over a period of time to give separate estimates of  $\lambda$ , say  $\lambda_j$ , for each person. These  $\lambda_j$ 's can then be used as dependent variables in a causal analysis by relating their variation to characteristics of the respondents or their situation.

Relying on counts of events is, however, often an inefficient use of the information available in life history data and may in fact provide misleading inferences. The basic assumption of the Poisson process is that events occur with a constant probability in each interval of time. Counts will have to take place over a time period, and with infrequent events this period may be quite long. It is likely that the causal variables relevant to the occurrence of events change over this period. This information is ignored when relying on counts. In other words, intra-individual variation cannot be studied when counts of events are used to study rates. Furthermore, the over-time variation in rates means that the counts do not estimate means in Poisson distributions, so what is studied is not well defined.

An example that illustrates this point occurred in an analysis of job shifts that I did some years ago. One reasonable hypothesis about the occurrence of job shifts is that they are more likely to occur the larger the discrepancy between a person's occupational resources (education, ability, etc.) and the returns obtained in the job in the form of status and earnings. Such a hypothesis cannot be tested using counts of events to estimate the rate of shifts, since the returns a person obtains from jobs will change over time as a result of the very job shifts that are analyzed. A different approach is needed, and it is offered by relying on waiting times.

The rationale used for waiting times is that if the occurrence of events is Poisson, waiting times will be exponentially distributed with mean  $1/\lambda$ . The assumptions of course are the same as for the Poisson distribution. However, waiting times need not be summed over time as in the case of counts of events. Rather each waiting

time may be treated as a unit of analysis. This means that if there are  $N$  individuals in the sample and  $k$  events for each individual, there will be  $N \cdot k$  units of observation available for analysis. Each configuration of values of the independent variables may be seen as defining a different Poisson process with its associated exponential distribution, and the procedure of treating waiting times as units will provide a set of means for these processes. The procedure thus provides meaningful quantities also with within-individual variation.

In the analysis of job shift each duration of a job was treated as an observation of the dependent variable, and this variable was then analyzed for its dependency on variables characterizing individuals and their jobs. The aforementioned hypothesis was substantiated. Straightforward OLS regression was used. This was probably not the best choice of estimation technique. A maximum likelihood procedure has been developed by Tuma that has more desirable statistical properties and also permits the use of independent variables, such as age, that vary continuously over the period of observation<sup>11</sup>.

The proposed procedure is then to use observations on intervals of time between events to estimate expressions of the form:

$$\lambda = b_0 + \sum_i b_i x_i \quad (19)$$

and to use estimates of the  $b_i$  coefficients to make inferences about the causes of variation in the occurrence of events. The linear specification may seem a convenient choice. There is, however, one important reason for choosing a different specification. What is analyzed are rates, and they are non-negative quantities. Hence, for example:

$$\lambda = \exp(b_0 + \sum_i b_i x_i) \quad (20)$$

may be a better choice.

The use of waiting times gives rise to a rather intriguing problem. It will usually be the case that observations are terminated at an arbitrary point in time in relation to the process. This means that the last waiting time until an event will be interrupted by, for example, the interview. The problem is what to do with this interval. It can be shown that if all other intervals of time are exponentially distributed the truncated interval will be gamma distributed with a mean that is twice that of other intervals. Intuitively the reason for this surprising result is that longer intervals of time have a greater chance of capturing the interruption than shorter intervals. The problem does affect estimation but several solutions are available<sup>12</sup>. It is, incidentally, not a solution to discard the truncated intervals, as serious bias may result.

<sup>11</sup> Tuma, N. B., Rewards, Resources and the Rate of Mobility: a Non-Stationary Multivariate Stochastic Model, in: *American Sociological Review*, 39 (1976), pp. 338–360.

<sup>12</sup> Sørensen, Aage B., Estimating Rates from Retrospective Data, in: Heise, D. (ed.), *Sociological Methodology*, 1977, San Francisco 1977.

## 7. Analyzing Outcomes of Events

The conditional probabilities of moving from state to state on the discrete variable given that an event occurs, the  $m_{ij}$ 's, may also be subjected to causal analysis. They can be estimated from event history data by counting the number of moves from each state of origin to each state of departure on each event. Thus, in an analysis of occupational mobility using event history data each job shift will result on a move from occupational category  $i$  to category  $j$ , where  $i$  may equal  $j$ . The unit of analysis is the shift. If there are  $N$  respondents and  $k$  shifts, there will be  $N \times (k-1)$  shifts available for analysis of the variation in the  $m_{ij}$ 's. The timing of shifts, and events in general, is of course irrelevant — it is analyzed using the approach described above.

The  $m_{ij}$ 's may be analyzed using an approach proposed by Spilerman<sup>13</sup>. For each row and cell in the  $m_{ij}$  matrix a variable  $y_{ij}$  is defined so that  $y_{ij} = 1$  if there is an entry in the  $ij$ 'th cell and  $y_{ij} = 0$  otherwise. For those outcomes originating in the  $i$ 'th row a regression analysis with  $y_{ij}$  as the dependent variable is performed, i. e. the expression:

$$y_{ij} = a_0 + \sum_i a_i x_i \quad (21)$$

is estimated. There will be  $k^2$  such equations with  $k$  states or categories of the discrete variable being analyzed.

Spilerman proposed the procedure for the analysis of transition probabilities in a discrete time Markov model, not for analysis of the  $m_{ij}$ 's. However, the discrete time transition probabilities estimated, for example, from panel data confounds the rate at which events occur with the outcome of events when they are estimated from a continuous time process. Though the abundance of panel data makes it tempting to treat event-history data with techniques developed for panel data the result is an inefficient use of the information contained in life histories. Direct analysis of the  $m_{ij}$ 's that govern the outcome of events is preferable.

Equation (21) is a linear probability model and the use of ordinary least squares is inefficient and the linear form is probably a misspecification. Log-linear analysis of the  $m_{ij}$ 's should be preferable.

An interesting parallel between the continuous variable and the discrete variable case should be noted. In the survey of models for change in continuous variables it was pointed out that the equilibrium state of the model for change with feedback is the simple linear model used in regression analysis of cross-sectional data to the relationship among variables. A similar result may be obtained for the discrete variable case, at least in the two state situation.

The Markov Chain will result in an equilibrium distribution if certain restrictions on the transition rates are fulfilled (corresponding to the condition  $b < 0$  for equa-

<sup>13</sup> Spilerman, S., The Analysis of Mobility Processes by the Introduction of Independent Variables in a Markov Chain, in: *American Sociological Review*, 37 (1974), pp. 277–294.



tion (7) to reach an equilibrium state). The equilibrium distribution will reflect the  $m_{ij}$ 's as the rate at which events occur will determine only the speed with which equilibrium is reached. In the two state case the equilibrium distribution will be the vector  $p^{(\infty)}$  with elements  $p_1^{(\infty)}$  and  $p_2^{(\infty)}$ . In terms of the  $m_{ij}$ 's these two quantities can be written as:

$$p_1^{(\infty)} = \frac{m_{21}}{m_{12} + m_{21}}$$

and

$$p_2^{(\infty)} = \frac{m_{12}}{m_{12} + m_{21}} \quad (22)$$

Now, let the  $m_{ij}$ 's be log-linear functions of independent variables, that is:

$$m_{12} = \exp(b_o + \sum_i b_i x_i)$$

$$m_{21} = \exp(c_o + \sum_i c_i x_i) \quad (23)$$

It follows, inserting (23) into (22) and taking the ratio of  $p_1^{(\infty)}$  and  $p_2^{(\infty)}$

$$\frac{p_1^{(\infty)}}{p_2^{(\infty)}} = \exp[b_o - c_o + \sum_i (b_i - c_i)x_i] \quad (24)$$

or

$$\log \frac{p_1^{(\infty)}}{p_2^{(\infty)}} = (b_o - c_o) + \sum_i (b_i - c_i)x_i \quad (25)$$

Equation (25) is the usual form of the logit model, and if the  $x_i$  variables are dummy variables then it is just a special case of Goodman's (1972) log-linear model for odds-ratios. Hence the log-linear model for odds-ratios may be seen as the equilibrium formulation of the Markov Chain model for change in discrete variables with an exponential decomposition of the  $m_{ij}$ 's in terms of independent variables. The proof for the two-state case has previously been given by Tuma, Hannan and Groenfeld, who, however, rely on the transition rates, the  $q_{ij}$ 's. of equation (11), rather than the  $m_{ij}$ 's<sup>14</sup>. If the  $m_{ij}$ 's are written as linear functions of independent variables, it can be shown — slightly modifying an approach suggested by Coleman<sup>15</sup> — that the linear probability model results.

As in the case of continuous variables the *ad hoc* statistical models that may be used to establish the relationships among discrete variables can be seen as equilibrium states of the simplest models for change. It follows conversely, that if the  $m_{ij}$ 's are being subject to log-linear analysis, it is assumed that the  $m_{ij}$ 's are in equilibrium.

<sup>14</sup> Tuma, N. B., et al., Dynamic Analysis of Social Experiments, paper presented at the 1977 meetings of the American Sociological Association.

<sup>15</sup> Coleman, Introduction.

librium. This assumption is, however, usually more realistic than assuming that the state distribution is in equilibrium. For example, in analysis of occupational mobility the occupational distribution of a cohort will usually change with the age of the cohort as individuals form their occupational careers. Stable  $m_{ij}$ 's are consistent with such an outcome. These quantities govern the outcome of moves when they occur and may be assumed to reflect the occupational structure and be quite stable, while the rate of movement changes with age.

## 8. Conclusion

This paper has advocated an approach to the analysis of change in discrete or categorical variables where stochastic process models are used to identify the components of change and causal analysis of the sources of variation in these components is then carried out. The continuous time Markov Chain has been suggested as the appropriate framework for such analysis. With event-history data this framework can be utilized to analyze the rate at which events occur and outcomes of events as functions of variables assumed to be relevant to change processes.

As mentioned above, the Markov Chain is usually not able to predict the course of observed social processes very adequately. It may seem that choosing this model as a framework is an unfortunate choice. However, the failure of the model is often due to failure of the assumptions of stationarity and homogeneity. The analysis proposed here are directed at identifying and accounting for variation in parameters over time and among individuals, and thus remedy these problems with the Markov model. The choice of this model is in fact not any more unrealistic than choosing the simple model for change with feedback as the framework for causal analysis of change in continuous variables, and this model has the linear equation, used so often in causal analysis, as its equilibrium state. It has been shown that the Markov model similarly has well known statistical models for analysis of relations among discrete variables as equilibrium formulations.

The alternative to the approach here is to use the *ad hoc* statistical techniques on change data and treat the over-time variation in the same manner as the cross-sectional variation. This approach has merit, but if the appropriate data on change are available — event-history data — these techniques do not make efficient use of the available information on change. Event-history data permit the direct analysis of change, and a framework that identifies the components of change is needed to take advantage of this opportunity.

# Have There Been Differences Between the Growth Rates in Different Periods of the Development of the Capitalist World Economy Since 1850?

## An Application of Cluster Analysis in Time Series Analysis

Cluster analysis is a method that can be employed to classify objects with different features in such a way that very similar objects are combined in one class (cluster), whereas dissimilar objects are combined in other classes (clusters). From a mathematical point of view, the objects are points in a multidimensional space, and the dimension of the space is given by the number of features. In pattern recognition terminology, the points are called patterns or OTU's (operational taxonomic units), the axes of the pattern space are the features or measurements.

Many different methods of cluster analysis are now in use, and the results of alternative methods are often not comparable. Hence the majority of mathematicians regard cluster analysis as a subjective technique, the application and interpretation of which depend on the user's standpoint, experience, and perspicacity. Nevertheless we hope to suggest an easy procedure for clustering time series that is not only usable but is also interpretable in a more objective way.

We can formulate the aim of the application of cluster analysis in terms of two requirements:

1. The variations within clusters should be as slight as possible.
2. The variations between clusters should be as large as possible.

For a mathematical formulation of the problem we introduce the following symbols and definitions:

$t = 1, 2, \dots, N$  — index of objects

$i = 1, 2, \dots, m$  — index of clusters

$j = 1, 2, \dots, n_i$  — index of objects within the  $i$ -th cluster

$k = 1, 2, \dots, p$  — index of features

$x_{ijk}$  —  $k$ -th feature of the  $j$ -th element within the  $i$ -th cluster

$\bar{x}_{ik} = \sum x_{ijk}/n_i$  —  $k$ -th feature of the  $i$ -th cluster-mean

$\bar{x}_k = \sum \bar{x}_{ik}/m$  — (unweighted) mean of the cluster-means of the  $k$ -th feature

We now have to answer the question of how to measure the variations. The most well-known formulation for the measurement of variations is the variance, and

*Acknowledgement:* For helpful assistance in programming I have to thank Dorothea Holland (Centre for Computing Techniques of the Academy of Sciences of the GDR).

many cluster methods are based on the computation of within- and between-cluster variances. But we must not forget that in this way we would compute the measure in squared units. For instance the objective function of a cluster analysis of population data would be formulated in „squared persons“, and that is – from every non-mathematical point of view – nonsense. A second possibility of variation measurement is the square root of the variance, the standard deviation. The disadvantage of the standard deviation is that it depends upon magnitude of the features. In respect to different features this disadvantage could be eliminated by standardization. But in this way the problem is not soluble in respect to one feature, since the absolute difference between two units of one feature still depends upon their size. Larger values, mostly, involve larger differences, and vice versa. Therefore, we should use the coefficient of variation (V).

$$V(x_{ik}) = \sqrt{\frac{1}{n_i - 1} \sum_j (x_{ijk}/x_{ik} - 1)^2} \quad (1)$$

The within-cluster variation in respect to the k-th feature is measurable as the weighted mean of the  $V(x_{ik})$ , i. e.

$$V_k^{(w)} = \frac{1}{N} \sum_i n_i V(x_{ik}) \quad (2)$$

Analogous to formula (1), we compute the between-cluster variation:

$$V_k^{(b)} = \sqrt{\frac{1}{m-1} \sum_i (x_{ik}/x_k - 1)^2} \quad (3)$$

The two requirements indicated above can be formulated for each feature as extremum problems:

$$V_k^{(w)} \stackrel{!}{=} \min \quad (4a)$$

$$V_k^{(b)} \stackrel{!}{=} \max \quad (4b)$$

Of course, the difference  $V_k^{(b)} - V_k^{(w)}$  is the objective function that must be maximized for each feature:

$$c_k = V_k^{(b)} - V_k^{(w)} \stackrel{!}{=} \max \quad (5a)$$

Consequently, the objective function of the whole system is:

$$C = \sum_k g_k c_k \stackrel{!}{=} \max \quad (\text{with } \sum_k g_k = 1) \quad (5b)$$

where  $g_k$  are the weights of the features (if we compute C without weights then each  $g_k$  equals  $1/p$ ).

In general in cluster analysis, the optima of the objective functions C are not computable because the number of possible solutions is  $(N+1)^N$ , and no computer

has a memory large enough to solve the problem. Hence we must introduce restrictions and/or a priori informations, e. g. a given number of clusters or a given minimum of within-cluster variation, or we used so called heuristic procedures, and in that case we again land at the starting point and come face to face with „the user's dilemma“<sup>1</sup>.

But I consider that all those heuristic procedures, restrictions, a priori information etc. are not necessary in order to carry out a time series analysis by means of cluster analysis. Firstly we do not use a method based on a minimization of squared errors — like FORGY, ISODATE, WISH etc. — that would force us to give a minimum<sup>2</sup>. Secondly, a time series is a well-ordered set, the sequence of data is given — even as a time series. The problem whether or not the  $t$ -th and the  $(t+v)$ -th year are elements of the same, let us say the  $i$ -th period, is soluble by means of stepwise clustering. If we state for the  $(t+u)$ -th year ( $u < v$ ) that is not an element of the  $i$ -th period, then, of course, neither is the  $(t+v)$ -th year an element of that period. Hence, and thirdly, we are not forced to specify the number of clusters (like in CLUSTER). We can use an agglomerative method of cluster analysis for the *mathematical* solution of periodization problems.

The criterion for this periodization is a purely mathematical one, i. e. the subjective option of both the historian and the mathematician is eliminated, provided, however, that we accept this criterion as such. Acceptance of the criterion however, depends on whether or not it reflects the mathematical structure of the object under investigation. But we can only answer this question from a historical point of view, the basis for the applicability of a mathematical method is to be found in the object under investigation, and is not to be found in the mathematical method itself. It is absolutely necessary to stress this fact because the application of mathematical methods in any field of science without a scientific that is a non-mathematical foundation is only nonsense. Quite rightly, Hegel spoke about the mathematical conclusion as the exterior conclusion („äußerer Schluß“)<sup>3</sup>.

Moreover, clustering techniques are tools for discovery rather than ends in themselves and should permit the user form statistical questions for further studies<sup>4</sup>. In the case of time series analysis, one such question would concern the significance of the difference between the within-periods-means.

As an example we analyze the growth of production and trade (in terms of constant prices) in the capitalist world since 1850. Figures 1—7 show us the growth of

<sup>1</sup> Dubes, R., and Jain, A. K., Clustering Techniques: The User's Dilemma, in: Pattern Recognition, 8 (1976), pp. 247 ff.

<sup>2</sup> Otherwise we would only obtain the trivial solution: Only years with identical features would constitute a cluster.

<sup>3</sup> Hegel, Encyclopedia of Philosophical Science (1830), § 188 (v. Wissenschaft der Logik, Vol. III, ch. 1.3.A.d.).

<sup>4</sup> Dubes, Techniques.

- |                                    |   |                                    |
|------------------------------------|---|------------------------------------|
| – industrial production            | } | in Million Dollars                 |
| – agricultural production          |   |                                    |
| – total production (incl. mining)  |   |                                    |
| – total exports                    |   |                                    |
| – ratio of industrial production   | } | to total production<br>in promille |
| – ratio of agricultural production |   |                                    |
| – ratio of total exports           |   |                                    |

By means of regression analysis we can demonstrate that the growth of production and trade over the past 125 years was exponential<sup>5</sup>. The quotient of two exponentially growing indicators also grows exponentially. Consequently, we can say that the capitalist world economy is an expanding economy. Therefore it is useless to cluster the series of states (as given in Appendix I)<sup>6</sup>, for the result of such an attempt is a very large number of clusters with a very small number of elements, even for years with similar states. For example the result of such a cluster analysis of the industrial production of states is division of this time series into 33 clusters. This fact only shows that the mathematical structure of this time series is not such that we can use the method in question. The result of the application of the method shows that the method is not applicable.

Because of the exponential growth of the capitalist world economy, we compute a time series of growth rates; symbolizing the quanta produced or sold by  $x_t$ , we symbolize the indices by

$$y_t = 100 x_t / x_{t-1} \quad (6)$$

Then the theoretical values are

$$\hat{x} = ab^t \quad (7a)$$

$$\hat{y} = 100 \hat{x}_t / \hat{x}_{t-1} = 100 ab^t / ab^{t-1} = 100 b \quad (7b)$$

The empirical indices  $y_t$  oscillate around their theoretical values  $100 b$ . Therefore it is useful to analyze indices. As to our mathematical apparatus, it would be a mistake to apply it to the indices  $y_t$  because the method of least squares cannot be employed for the analysis of quotients. This follows from the fact that the arithmetic mean of quotients does not equal the quotient of arithmetic means:

$$\frac{100}{n} \sum x_t / x_{t-1} \neq \frac{\frac{100}{n} \sum x_t}{\frac{1}{n} \sum x_{t-1}} \quad (8a)$$

<sup>5</sup> Kuczynski, Th., Spectral Analysis and Cluster Analysis as Mathematical Methods for the Periodization of Historical Processes – a Comparison of Results Based on Data about the Development of Production and Innovation in the History of Capitalism. Kondratieff Cycles – Appearance or Reality, in: Proceedings of the Seventh International Economic History Congress, Edinburgh 1978, pp. 79.

<sup>6</sup> Sources: Kuczynski, J., Die Geschichte der Lage der Arbeiter unter dem Kapitalismus, Vol. 37, Berlin 1967, p. 31, 78 (1850–1964); UNO-Monthly Bulletin of Statistics (1964–1976).

Therefore we use the logarithmic indices

$$z_t = 2 + \lg x_t - \lg x_{t-1} \quad (9a)$$

which oscillate around the expected value

$$\hat{z}_t = 2 + \lg b \quad (9b)$$

The arithmetic mean of the logarithmic indices  $M_a(z)$  equals the logarithm of the geometric mean of the indices  $\lg(M_g(y))$ :

$$\begin{aligned} M_a(z) &= \frac{1}{n} \sum (2 + \lg x_t - \lg x_{t-1}) \\ &= 2 + \frac{1}{n} (\lg x_n - \lg x_0) \\ &= \lg(100 \sqrt[n]{x_n/x_0}) \\ &= \lg(\sqrt[n]{100x_t/x_{t-1}}) \end{aligned} \quad (10a)$$

hence

$$M_a(\lg y) = \lg(M_g(y)) \quad (10b)$$

Table 1 shows the logarithmic growth rates

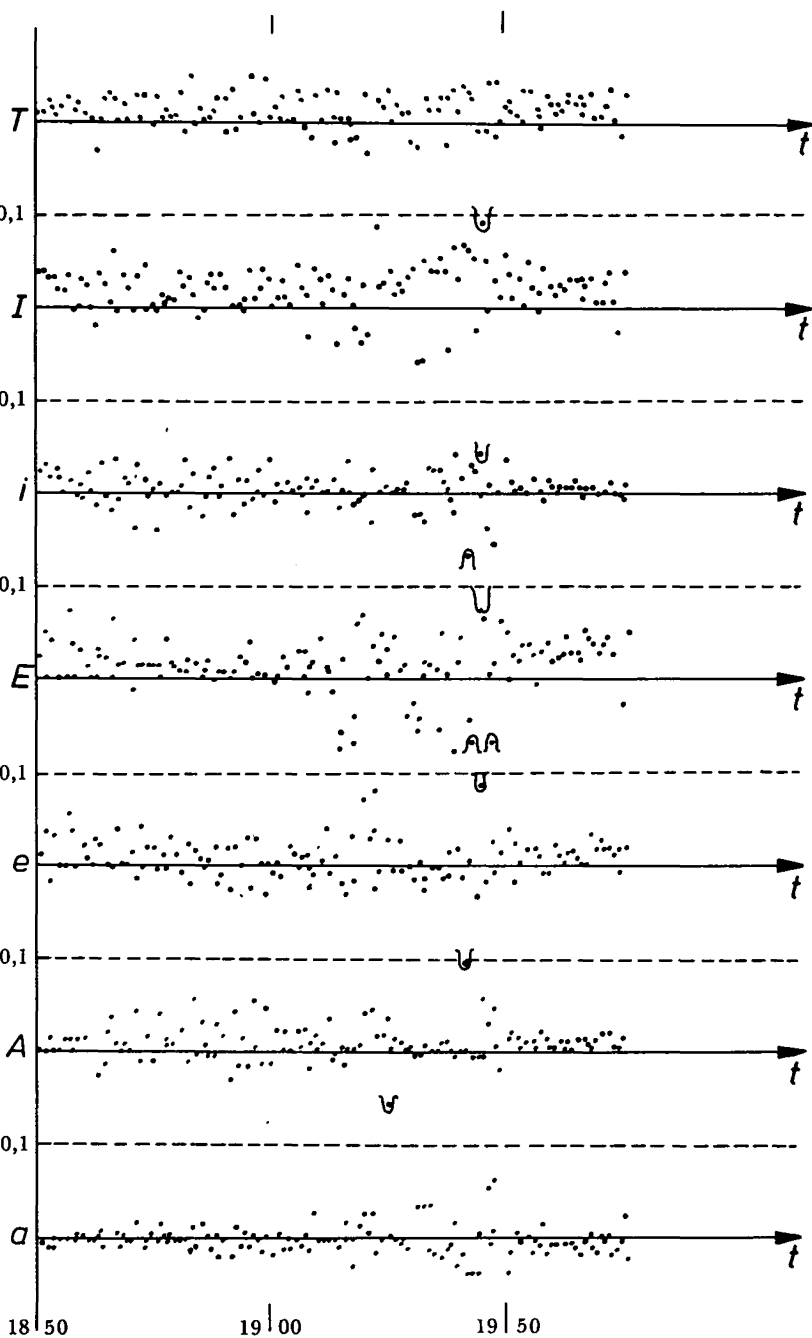
- (1) agricultural production (A)
- (2) industrial production (I)
- (3) total production (including mining) (T)
- (4) total exports (E)
- (5) the ratio of agricultural to total production (a)
- (6) the ratio of industrial to total production (i)
- (7) the ratio of total exports to total production (e)

We can now ask a more correct question: Are there differences between the logarithmic indices in different periods of development of capitalist world economy since 1850? We want to try to answer this question by using the clustering technique described above.

If we cluster the six-dimensional time series (without (4)) the optimum result is given by a division of these series into two clusters: 1851–1945, 1946–1976. We get a local optimum if we construct the clusters 1850–1920, 1921, 1922, 1923–1942, 1943, 1944, 1945, 1946, 1947–1976. The results of the one-dimensional cluster analysis of all seven time series are similar. Because of the world wars, which most strongly affected the (logarithmic) indices, the results are almost useless.

Because there is such great irregularity in the indices between the world wars, we build an a priori cluster and repeat the cluster analysis. The results are completely different. Before World War I, we can differentiate three periods in all seven series;

Table 1: Growth Rates of the Capitalist World Economy 1850–1976 (logarithmic scale 1 mm  $\cong$  0,0050/1 year  $\cong$  1 mm)





after World War II we can differentiate two periods in four of the seven series (in the others we cannot identify different periods). On the other hand, the six-dimensional cluster analysis shows us only one pre-war period and one post-war period, a fact which is caused by the shift of the single series.

But we have to ask ourselves whether the differences between the average growth rates are random or significant. As shown in formula (10a), these averages depend only on the last year of the given period ( $z_{in_i}$ ) and the year before the first year of the given period ( $z_{i0} = z_{i-1, n_{i-1}}$ ). Although the within-cluster differences are considered in formula (5), it is useful to test the differences between the average growth rates.

First, we test whether the data are normally distributed. By means of the Kuiper test<sup>7</sup> we prove that the logarithms of the indices are normally distributed, i. e. the indices are log-normally distributed. If the logarithmic indices are normally distributed, then the logarithmic growth rates

$$q_t = z_t - 2 \quad (11a)$$

are also normally distributed. The variance of the logarithmic growth rates is of the same magnitude as the variance of the logarithmic indices:

$$D^2(q) = D^2(z) \quad (11b)$$

Because the variances of  $q$  or  $z$  are different in different periods we test the significance of the differences between two average (logarithmic) growth rates by means of the approximate test by Welch<sup>8</sup>. The data needed for computing the test variable  $t$  and the degree of freedom  $f$  are just the same as for computing the objective function (5a). The average differ significantly only if

$$t = \frac{x_{ik} - x_{i-1, k}}{\sqrt{V^2(x_{ik}) x_{ik}^2 / n_i - V^2(x_{i-1, k}) x_{i-1, k}^2 / n_{i-1}}} \rightarrow t_{f, \alpha} \quad (12a)$$

with

$$f = \frac{(n_i - 1)(n_{i-1} - 1)(V_{ik}^2 x_{ik}^2 n_{i-1} + V_{i-1, k}^2 x_{i-1, k}^2 n_i)^2}{V_{ik}^4 x_{ik}^4 n_{i-1}^2 (n_{i-1} - 1) + V_{i-1, k}^4 x_{i-1, k}^4 n_i^2 (n_i - 1)} \quad (12b)$$

and  $\alpha$  as the significance level (95 %).

<sup>7</sup> Kuiper, N. H., Tests Concerning Random Points on a Circle, in: Proc. Konink. Ned. Akad. Wet., A 63 (1960), pp. 38 ff.

<sup>8</sup> Welch, B. L., The Significance of the Difference Between Two Means When the Population Variances are Unequal, in: Biometrika, 29 (1938), pp. 350 ff.

Because of the fact that we use an agglomerative method for the clustering of time series, we can not directly combine the Welsh test with this method. Enumerating the solutions of the objective functions  $c_k$  or  $C$  according to the number of clusters, we first have to compute the solutions

$c_k^{(N)}, c_k^{(N-1)}, \dots, c_k^{(1)}, \dots, c_k^{(2)}, c_k^{(1)}$ . We then carry out the paired comparison of the averages  $x_{ik}^{(1)}$  and  $x_{i-1,k}^{(1)}$ . The figure of this desagglomeration (or division) is the inverted dendrogramme which reflects the process of agglomeration. The results of cluster analysis and Welsh test are summarized in Appendix II.

We can say there have been periods of slower growth and periods of faster growth in respect to both industrial production and the ratio of industrial to total production. But this change is not statistically significant. If we assume a level of confidence probability of 95 %, then we find only one pair of logarithmic indices (or logarithmic growth rates) which differ significantly. All the other  $z_i$  differ significantly only at a lower level of confidence probability, that lies between 75 % and 90 %.

From a scientific standpoint we can only say that there have been differences between the cluster-means but the differences are too slight (or the variances of the means are too large) to be considered significant. Speaking positively, such a statement is an invitation for further research.

The results of clustering and testing export series are completely different. We are able to discriminate clearly three periods: 1850/66, 1867/1949, 1950/1976 the means of which differ significantly. The subperiods 1867/1893, 1894/1913 and 1913/1949 are very similar to each other. The growth of agricultural production and of its ratio to total production is relatively regular: The cluster means differ very little (whereas the coefficients of variation are similar to those of the other series).

The answer to our original question is not very satisfying: There have been differences but most of them are not significant. The main reason for this is that variances within the clusters are very large. Most of the approximate coefficients of variation of the growth rates (i. e.  $v(r_i)$ ) are greater than one hundred per cent. Moreover, we are able to prove that the size of  $v(r_i)$  depends in such a way on  $r_i$  that the smaller the absolute value of  $r_i$  the larger its coefficient of variation. We construct a two-by-two table to show this fact. For that purpose we compare  $r_{ik}$  and  $v(r_{ik})$  with  $r_k$  and  $v(r_k)$  respectively. The table shows that high values of  $v(r_{ik})$  are related to low values of  $r_{ik}$ :

	$r_{ik} > r_k$	$r_{ik} < r_k$
$v(r_{ik}) > v(r_k)$	1	9
$v(r_{ik}) < v(r_k)$	16	7

By means of Chi-squared-test (Yates correction included) we show that the probability of the mentioned dependence is more than 99 % (by means of the Fisher-Yates test we obtain a similar result: The rejection probability for our hypothesis is only 0.21 %) <sup>9</sup>.

So we can conclude that the slower the growth of capitalist world economy the stronger the irregularities affecting the growth rates. Therefore it is not very useful to treat the time before World War I as a homogenous period. Quite clearly, there was a break in the development which relates to the transition of capitalist world economy to its imperialist stage. This break is visible in the series on industrial production and its ratio to total production. After the transition period, at the end of the last century, growth rates began to increase again. But that did not occur in the field of foreign trade. The growth rates of export volume increased slightly, but the international division of labor the ratio of total exports to total production hardly changed. Protectionism continued, and only after World War II did a strong up-swing period set in.

Looking at the period after World War II it seems very probable that in the second half of the sixties a long down-swing period began. Too short a time has gone by to decide this question definitively but some indicators give support to such an opinion. So we may assume that the growth of the capitalist world economy over a longer period will be slower and more unstable than in the first twenty years after World War II.

<sup>9</sup> See e. g. Fisher, R. A., *Statistical Methods for Research Workers*, London 1948, p.95.

# Appendix I: The Growth of Capitalist World Economy 1850-1976

Figure 1: Industrial Production  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	31	34	37	40	42	44	48	47	47	50
186.	53	53	47	51	54	55	63	62	67	70
187.	68	74	82	81	81	79	82	83	85	87
188.	94	99	106	110	107	106	112	121	126	137
189.	144	146	148	143	148	162	166	174	190	204
190.	205	213	230	235	236	260	272	280	257	282
191.	302	307	334	350	319	329	354	357	336	308
192.	326	280	343	361	382	413	427	455	476	511
193.	448	389	333	375	417	466	525	578	518	606
194.	651	777	907	1033	980	679	578	651	700	714
195.	812	882	903	970	970	1085	1138	1176	1145	1274
196.	1358	1407	1500	1576	1705	1831	1975	2004	2130	2288
197.	2335	2382	2568	2802	2825	2638	2895			

Figure 2: Agricultural Production  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	116	116	120	120	120	124	128	128	132	136
186.	136	136	128	124	132	144	148	148	152	152
187.	168	164	172	172	168	184	184	192	196	200
188.	212	196	224	216	232	236	232	228	244	256
189.	236	264	252	260	252	288	292	280	312	300
190.	316	312	332	328	328	348	368	348	356	372
191.	372	376	408	400	388	392	352	352	352	360
192.	396	396	440	456	472	512	512	528	544	544
193.	548	544	540	548	548	548	564	592	588	576
194.	576	576	572	552	536	460	524	560	624	632
195.	596	612	640	664	668	688	708	708	740	764
196.	780	788	808	820	844	848	880	912	936	944
197.	964	996	996	1044	1060	1080	1118			

*Figure 3: Total Exports*  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	17	19	19	21	21	21	25	27	27	29
186.	30	30	30	32	34	36	42	42	44	46
187.	44	49	51	53	55	57	59	59	65	68
188.	70	72	76	78	82	82	84	89	89	91
189.	93	93	95	95	101	106	116	116	118	120
190.	118	120	120	127	131	141	152	154	146	152
191.	160	177	184	190	158	137	144	125	114	131
192.	152	152	165	173	194	207	209	232	238	247
193.	226	211	184	186	194	201	207	232	213	213
194.	174	183	202	262	232	158	188	220	224	236
195.	273	308	304	323	338	367	399	422	410	441
196.	485	509	538	578	636	685	736	769	875	970
197.	1054	1128	1233	1381	1465	1370	1539			

*Figure 4: Total Production Including Mining*  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	149	153	160	163	165	171	180	179	183	190
186.	193	193	179	180	191	204	216	216	225	228
187.	242	245	262	261	257	271	274	283	290	296
188.	316	306	342	339	351	354	357	363	385	408
189.	396	427	417	420	417	469	478	475	525	528
190.	547	552	590	593	594	642	677	667	649	694
191.	717	727	789	800	750	765	757	762	740	712
192.	773	716	832	874	911	986	1000	1050	1089	1130
193.	1065	990	921	973	1021	1076	1160	1250	1180	1259
194.	1310	1438	1566	1675	1604	1212	1182	1299	1419	1439
195.	1507	1602	1654	1748	1751	1897	1978	2021	2015	2175
196.	2284	2348	2469	2562	2725	2861	3028	3114	3275	3459
197.	3536	3618	3790	4112	4181	3996	4281			

*Figure 5: Ratio Industrial to Total Production (Promille)*  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	208	223	231	245	255	257	267	263	257	263
186.	275	275	263	283	283	270	292	287	298	307
187.	281	302	313	310	315	291	299	293	293	294
188.	297	324	310	324	305	299	314	333	327	336
189.	364	342	355	343	355	345	347	366	362	386
190.	375	386	390	396	397	405	402	418	396	406
191.	421	422	423	438	425	430	468	468	454	433
192.	422	391	412	413	419	419	427	433	437	452
193.	421	393	361	385	408	433	453	462	439	481
194.	497	540	579	617	611	560	489	501	493	496
195.	539	551	546	555	554	572	575	582	568	586
196.	595	599	608	615	626	640	646	664	650	661
197.	660	658	678	681	676	660	676			

*Figure 6: Ratio Agricultural to Total Production (Promille)*  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	779	758	750	736	727	725	711	715	721	716
186.	705	705	715	689	691	706	685	685	675	667
187.	694	669	656	659	654	679	671	678	676	676
188.	671	641	655	637	661	667	650	628	634	627
189.	596	618	604	619	604	614	611	590	594	568
190.	578	565	563	553	552	542	539	519	549	536
191.	519	517	517	500	517	512	465	451	476	505
192.	512	553	529	522	518	519	512	503	500	481
193.	515	550	586	563	537	509	486	474	498	458
194.	440	400	365	330	334	380	443	431	440	439
195.	395	382	387	380	382	363	358	350	367	351
196.	342	336	327	320	310	296	291	300	286	273
197.	273	275	263	254	254	270	261			

**Figure 7: Ratio Total Exports to Total Production (promille)**  
(100 Mill. Dollars in constant prices of 1913)

Year	... 0	... 1	... 2	... 3	... 4	... 5	... 6	... 7	... 8	... 9
185.	114	124	119	129	127	123	139	151	148	153
186.	155	155	168	178	178	176	194	194	196	202
187.	182	200	195	203	214	210	215	208	224	230
188.	221	235	222	230	234	232	235	245	231	223
189.	235	218	228	226	242	226	243	244	225	227
190.	216	217	203	214	221	220	224	230	225	219
191.	223	244	233	238	211	179	190	164	154	184
192.	197	212	198	198	213	210	209	221	219	219
193.	212	213	200	191	190	187	178	186	181	169
194.	133	127	129	156	145	130	159	169	158	164
195.	181	192	184	185	193	193	202	209	209	203
196.	212	217	218	226	233	239	243	247	267	280
197.	298	312	325	336	350	343	359			

## *Appendix II: Results of Cluster Analysis and Welch Test*

- $z_i$  — arithmetic mean of the logarithmic indices within the  $i$ -th cluster
- $D(z_i)$  — standard deviation of  $z_i$
- $V(z_i)$  — coefficient of variation of  $z_i$  (in %)
- $r_i$  — geometric mean of the growth rates within the  $i$ -th cluster =  $10^{(z_i - 2)}$  (in %)
- $v(r_i)$  — approximate coefficient of variation of  $r_i$  (in %) =  $D(z_i)/(z_i - 2)$
- $\alpha$  — confidence probability (in %); we used a sequence 60 %, 70 %, 75 %, 80 %, 85 %, 90 %, 95 %, 97,5 %, 99 %, 99,5 %, 99,95 %; if  $\alpha$  is lower than 60 % we wrote 50 %.

In the lower cluster we combined all those clusters the means of which are lower than that of the whole time series (1850–1976), and in the upper cluster we combined the others.

## 1. Industrial Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850-1866	2.01924	0.02587	1.28 %	4.8 %	134 %	75 % 80 % 80 % 90 % 80 %
1867-1896	2.01403	0.01714	0.85 %	3.3 %	122 %	
1897-1913	2.01906	0.02039	1.01 %	4.5 %	107 %	
1914-1950	2.00988	0.05294	2.63 %	2.3 %	536 %	
1951-1969	2.02368	0.01363	0.67 %	5.6 %	58 %	
1970-1976	2.01459	0.02479	1.23 %	3.4 %	170 %	80 %
1850-1976	2.01564	0.03294	1.63 %	3.4 %	211 %	75 % 90 %
1850-1913	2.01671	0.02013	1.00 %	3.9 %	120 %	
1914-1950	2.00988	0.05294	2.63 %	2.3 %	536 %	
1951-1976	2.02123	0.01727	0.85 %	5.0 %	81 %	
upper cl.	2.02080	0.01970	0.97 %	4.9 %	95 %	90 %
lower cl.	2.01200	0.03942	1.96 %	2.8 %	328 %	

## 2. Agricultural Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850-1880	2.00873	0.01579	0.79 %	2.0 %	181 %	60 % 60 % 70 % 85 %
1881-1894	2.00536	0.02932	1.46 %	1.2 %	547 %	
1895-1915	2.00914	0.02193	1.09 %	2.1 %	240 %	
1916-1950	2.00520	0.02408	1.20 %	1.2 %	463 %	
1951-1976	2.01051	0.00835	0.42 %	2.4 %	79 %	
1850-1976	2.00781	0.02001	1.00 %	1.8 %	256 %	50 % 85 %
1850-1915	2.00814	0.02098	1.04 %	1.9 %	258 %	
1916-1950	2.00520	0.02408	1.20 %	1.2 %	463 %	
1951-1976	2.01051	0.00835	0.42 %	2.4 %	79 %	
upper cl.	2.00944	0.01566	0.78 %	2.2 %	166 %	80 %
lower cl.	2.00524	0.02537	1.27 %	1.2 %	484 %	



### 3. Total Exports

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850—1866	2.02455	0.02472	1.22 %	5.7 %	101 %	90 %
1867—1893	2.01313	0.01331	0.66 %	3.1 %	101 %	60 %
1894—1913	2.01506	0.01368	0.68 %	3.5 %	91 %	85 %
1914—1949	2.00261	0.05502	2.75 %	0.6 %	2108 %	99 %
1950—1976	2.03016	0.02032	1.00 %	7.1 %	67 %	
1850—1976	2.01553	0.03432	1.70 %	3.6 %	220 %	
1850—1913	2.01623	0.01567	0.78 %	3.8 %	97 %	90 %
1914—1949	2.00261	0.05502	2.75 %	0.6 %	2108 %	99 %
1950—1976	2.03016	0.02032	1.00 %	7.1 %	67 %	
upper cl.	2.02235	0.02226	1.10 %	5.3 %	100 %	99 %
lower cl.	2.00712	0.04255	2.12 %	1.7 %	598 %	

### 4. Total Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850—1866	2.01008	0.01483	0.74 %	2.3 %	147 %	60 %
1867—1893	2.00734	0.01687	0.84 %	1.7 %	230 %	90 %
1894—1913	2.01190	0.01978	0.98 %	2.8 %	166 %	85 %
1914—1950	2.00743	0.02650	1.32 %	1.7 %	357 %	97.5 %
1951—1966	2.01894	0.01293	0.64 %	4.5 %	68 %	75 %
1967—1976	2.01482	0.01657	0.82 %	3.5 %	112 %	
1850—1976	2.01050	0.01938	0.96 %	2.7 %	185 %	
1850—1913	2.00948	0.01728	0.86 %	2.2 %	182 %	
1914—1950	2.00743	0.02650	1.32 %	1.7 %	357 %	85 %
1951—1976	2.01736	0.01433	0.71 %	4.1 %	83 %	97.5 %
upper cl.	2.01465	0.01642	0.82 %	3.4 %	112 %	85 %
lower cl.	2.00825	0.04600	2.29 %	1.9 %	558 %	

### 5. Ratio Industrial to Total Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850-1869	2.00939	0.01653	0.82 %	2.2 %	176 %	95 % 75 % 90 % 95 % 85 %
1870-1896	2.00198	0.02055	1.03 %	0.5 %	1038 %	
1897-1913	2.00590	0.01295	0.65 %	1.4 %	219 %	
1914-1950	2.00245	0.02435	1.22 %	0.6 %	994 %	
1951-1966	2.00493	0.00606	0.30 %	1.1 %	123 %	
1967-1976	2.00197	0.00689	0.34 %	0.5 %	350 %	
1850-1976	2.00406	0.01826	0.91 %	0.9 %	450 %	90 % 99.5 %
1850-1913	2.00512	0.01757	0.88 %	1.2 %	343 %	
1914-1950	2.00245	0.02435	1.22 %	0.6 %	994 %	
1951-1976	2.00739	0.00643	0.32 %	0.9 %	170 %	
upper cl.	2.00670	0.01276	0.64 %	1.6 %	190 %	90 %
lower cl.	2.00221	0.02119	1.06 %	0.5 %	959 %	

### 6. Ratio Agricultural to Total Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_i)$	$\alpha$
1850-1880	1.99784	0.00801	0.40 %	-0.5 %	371 %	70 % 60 % 50 % 80 % 60 %
1881-1890	1.99486	0.01345	0.67 %	-1.2 %	262 %	
1891-1915	1.99732	0.01127	0.56 %	-0.6 %	430 %	
1916-1950	1.99679	0.02822	1.41 %	-0.7 %	879 %	
1951-1966	1.99164	0.00987	0.50 %	-1.9 %	118 %	
1967-1976	1.99436	0.01579	0.79 %	-1.3 %	280 %	50 % 80 %
1850-1976	1.99614	0.01740	0.87 %	-0.9 %	453 %	
1850-1915	1.99720	0.01014	0.51 %	-0.6 %	362 %	
1916-1950	1.99679	0.02822	1.41 %	-0.7 %	879 %	
1951-1976	1.99268	0.01215	0.61 %	-1.7 %	166 %	90 %
upper cl.	1.99329	0.00987	0.50 %	-1.5 %	147 %	
lower cl.	1.99729	0.02396	1.20 %	-0.6 %	884 %	

# 7. Ratio Total Exports to Total Production

Cluster	$z_i$	$D(z_i)$	$V(z_i)$	$r_i$	$v(r_0)$	$\alpha$
1850—1866	2.01447	0.02297	1.14 %	3.3 %	159 %	95 %
1857—1893	2.00244	0.02085	1.04 %	0.8 %	855 %	50 %
1894—1913	2.00106	0.02107	1.05 %	0.1 %	1988 %	75 %
1914—1949	1.99553	0.03775	1.89 %	-0.7 %	845 %	99 %
1950—1976	2.01263	0.01371	0.68 %	2.7 %	109 %	
1850—1976	2.00396	0.02676	1.34 %	0.9 %	676 %	
1850—1913	2.00506	0.02203	1.10 %	1.2 %	435 %	90 %
1914—1949	1.99553	0.03775	1.89 %	-0.7 %	845 %	99 %
1950—1976	2.01263	0.01371	0.68 %	2.7 %	109 %	
upper cl.	2.01331	0.01728	0.86 %	3.1 %	130 %	
lower cl.	1.99911	0.02925	1.46 %	-0.2 %	3287 %	99 %

## VI. Network Analysis



## Leadership Dimensions Among Developing Nations in the United Nations

### 1. Object of the Study and Theoretical Framework<sup>1</sup>

This paper contains an analysis of historical trends in group cohesion and leadership among developing nations within the United Nations General Assembly over the 1950's and 1960's. This study has several theoretical aspects.

For the developing nations the 1950's and 1960's were the transitional period from the colonial to the neo-colonial era, in which they created their own group- and negotiation structures to cope with their new situation. Within the United Nations the Afro-Asian countries started consultations in 1949 on an ad hoc basis; regular caucus meetings of this group began in 1955 after the Bandung Conference. In the 1950's Afro-Asian group formation was primarily focussed on problems of political independence and political non-alignment. Because of their remote interest in these problems, Latin American countries did not participate in these groups. Instead, they continued to hold their own regular caucus meetings as they had from the beginning of the United Nations. However, faced with the basic problems of economic dependence and development the politically independent Afro-Asian and Latin American countries began to realize that they had in the early sixties a common interest in a new policy for international trade and development. This led to the formation in 1963 of the „Group of 77“, in which they would coordinate their policies in the Second (economic) Committee of the General Assembly and the UNCTAD conferences.

*The fifth session (1950)* was chosen as our first year of analysis, because one can hardly speak of an Afro-Asian group before that year (Tadic, 1969). This group is

<sup>1</sup> This paper is one of the publications of a research project on legislative behavior at the Institute for Political Science of the University of Amsterdam. A more extensive publication on Third World group formation in the United Nations is forthcoming (Stokman, Frans N., Roll Calls and Sponsorship. A methodological analysis of Third World group formation in the United Nations, Leyden 1977). Prof. Dr. P. R. Baehr and Prof. Dr. R. J. Mokken of the Institute for Political Science of the University of Amsterdam were involved in this project on Third World group formation from its very beginning. In all phases of the study they provided valuable advice.

the most important subgroup we analyzed within the whole group of developing nations. By the time of the Korean war the Afro-Asian states realized that a war between the great powers might well be waged in Asia. This strongly stimulated Afro-Asian group formation and the predominantly Afro-Asian non-alignment movement.

*The 23rd session (1968)* was the last session studied, because the General Assembly Official Records of later sessions were not yet available in the Netherlands at the time the material for this study was collected. Given our research aims, the 23rd session is sufficiently recent, because the first years of cooperation of the developing nations in the „Group of 77“ are thus incorporated in our period of analysis.

The research period is divided into four periods:

- 5th – 10th session (1950–55)
- 11th – 14th session (1956–59)
- 15th – 18th session (1960–63)
- 20th – 23rd session (1965–68)<sup>2</sup>

We chose this particular division for several reasons. Within the different periods membership of the United Nations was relatively stable, whereas the differences between the periods are maximized in this respect. Moreover, the division roughly corresponds to generally observed and reported developments in group formation among the developing nations – in particular among the Afro-Asian nations – both inside and outside the United Nations.

At the end of the 10th session a large number of new members were able to enter the United Nations as the result of a „package deal“ between the United States and the Soviet Union in the Security Council. For the developments in the Afro-Asian group a split between 1954 and 1955 might have been more preferable. In the spring of 1955 Afro-Asian group formation was strongly stimulated at the Bandung Conference. From that time on, the Afro-Asian delegations at the United Nations met in regular caucus, the so-called Asian-African caucusing group.

At the beginning of the 15th session a large number of newly independent African States entered the United Nations en bloc. This drastically changed the relative strength of this group of states within both the Afro-Asian group and the United Nations as a whole. The Asian-African caucusing group was renamed Afro-Asian group and developments in Africa came to the fore.

The main reason for the division between the third and fourth period is the formation of the „Group of 77“ at the 1963 session. Can we discover differences in leadership and group cohesion before and after the formation of this group and before and after the experiences at the first UNCTAD-conference at Geneva in 1964?

<sup>2</sup> In the 19th session no votes were taken in the General Assembly in order to prevent suspension of the voting rights of the Soviet Union and France, which had fallen in arrears in their payments because of their refusal to contribute to the peacekeeping operations in the Congo and the Middle East. The session is therefore not included in our study.

A systematic analysis of leadership and group cohesion among the developing nations can be based on their behavior in the *UN General Assembly*, because the United Nations has always been an important instrument of their foreign policy. This has particularly been the case since 1960, when they achieved a majority position within the Assembly. This position has enabled them to transform their policy preferences into policy outputs of the United Nations, even when the great powers are opposed. This majority position had to be organized if it was to make effective use of the United Nations at all. Within this perspective, we can reformulate the main object of our study as a systematic study of the degree to which the developing nations acted as a group within the UN General Assembly; whether they did increasingly so over time and which delegations played an important role in this group process.

Such a systematic study of group processes within the UN General Assembly requires a *theoretical framework* in which the concepts of group, leadership and cohesion are related to the political process in the United Nations General Assembly. Only then are we able to determine the functions that the group must perform to realize its group goals.

In small group theories the concept *group* has a very broad meaning. Gibb, who reviewed research on leadership in small groups, defined a group on the basis of *interaction* of its members and *common purposes or goals*<sup>3</sup>. In conformity with this approach we shall speak of a *group of delegations*, if these delegations have *common purposes or goals and if they interact with each other to attain these goals*.

*Cohesiveness or cohesion of the group* refers to the degree to which the members of a group desire to remain in the group. It can be conceived as the resultant of all forces acting on members to remain in the group<sup>4</sup>. If we apply this definition to the United Nations, we can consider a group of delegations as cohesive to the extent that they agree on the group goals, the group is seen as a necessary instrument to reach the common goals and there is agreement with respect to the *group functions*, or activities that help the group achieve its group goals. *Leadership*, then, is viewed as the actual performance of these *group functions*.

Given the above definitions the first step is to identify the *group goals* of the developing nations at the time of our analysis. Socio-economic issues in the General Assembly can be considered as related to the group goals of the developing nations as a whole, because socio-economic problems have been the focus of the „Group of 77“ since its beginning. The important subgroup of Afro-Asian countries, however, has a broader value basis: colonial questions has been the other main issue area with which this group was concerned from the beginning. Our analysis has therefore been confined to the broad issue areas of *colonial and socio-economic problems*<sup>5</sup>.

<sup>3</sup> Gibb, Cecil A., Leadership, in: Lindzey, Gardner, and Aronson, Elliot (eds.), *The Handbook of Social Psychology*, vol. 4, 2nd ed., Reading/Mass. 1969, p. 207.

<sup>4</sup> Cartwright, Dorwin, and Zander, Alwin (eds.), *Group Dynamics Research and Theory*, 3rd ed., New York 1969, p. 91.

<sup>5</sup> Colonial and socio-economic questions were defined by enumeration of five subject cate-



Second, we must determine the *group functions* of the group of developing nations within the Assembly. The United Nations can be considered a political system, or more precisely, a political subsystem in the international political system, in which wants and demands are converted into authoritative allocations<sup>6</sup>. By group formation delegations are able to fix or change alternative allocations or to determine the outcomes of the allocation process within the system. In other words, the *primary function of group formation among developing nations is the exercise of political power and influence*. Political power refers to the capacity to fix or change (partly) a set of alternative allocations for the members of a political system or a part of it; political influence is the capacity to determine (partly) within a given set of alternative value allocations the outcomes of the allocation process<sup>7</sup>. However, political power and influence can be exercised in different phases of the political process<sup>8</sup>. Group formation, leadership and cohesion among developing nations can best be studied in that phase of the political process in which the developing nations try to exercise the most political power and influence. That is the decision-making phase. Thus, in their efforts to terminate colonial rule and to change their peripheral position in the international economic system, the developing nations have not been primarily oriented to prevention of decision-making in order to continue the status quo, but rather to a positive decision-making by the General Assembly. This has in particular been the case since 1960, because of their majority position within the Assembly. Since then they have had a very strong position in the decision-making phase of the political process in the General Assembly and as a consequence have directed their group efforts specifically to that phase and not to earlier phases.

Exercise of political power and influence in the decision-making phase implies that:

- 1) the developing nations should take *initiative* in the decision-making to formulate their policy aims and to bring them to the floor;

gories: (1) trust territories and the Trusteeship Council; (2) non-selfgoverning territories in Latin America (including the Caribbean), Asia, Africa and Oceania; (3) right of self determination; (4) problems between (ex-)colonial powers and former colonies; and (5) socio-economic problems (see Stokman, Roll calls).

<sup>6</sup> Politics and authoritative allocations are used here in a very broad sense. In this respect we follow De Vree, who adopted Easton's definition of politics (Easton, David, *A Systems Analysis of Political Life*, New York, 1965, p. 21) for use in the international setting (De Vree, J. D., *Political Integration: The Formation of Theory and Its Problems*, The Hague 1972, pp. 10 passim).

<sup>7</sup> Mokken, R. J., and Stokman, F. N., *Power and Influence as Political Phenomena*, in: Barr, Brian (ed.), *Power and Political Theory: Some European Perspectives*, New York 1976, p. 49; Helmers, H. M., et al., *Graven naar macht. Op zoek naar de kern van de Nederlandse economie* (Traces of Power, in Dutch), Amsterdam 1975, p. 65.

<sup>8</sup> Cf. particularly Bacharach, Peter, and Baratz, Morton S., *Power and Poverty. Theory and Practice*, New York 1970; Van der Eijk, C., and Kok, W. J. P., *Nondecisions Reconsidered*, in: *Acta Politica*, 10 (1975), pp. 277-301; Lukes, Steven, *Power. A Radical View*, London 1974.

2) the developing nations should maintain a sufficiently high level of *voting cohesion* to get their proposals adopted and to block undesired proposals.

*These two conditions can be considered the essential group functions of the developing nations and of important subgroups of developing nations within the General Assembly.* We can now ask whether the developing nations managed to satisfy these two conditions for the realization of their power position in the decision-making phase of the Assembly and which developing nations contributed most to this realization. In this paper we analyze only the performance of the first group function by the developing nations. For the analysis of the second group function and the relation between the two group functions the reader is referred to Stokman (1977 a).

*Leadership* will be indicated by *sponsorship* of resolutions, amendments and motions within the selected issue areas<sup>9</sup>. A group, if it is to change the status quo, must use resolutions to convert its policy preferences into United Nations decisions, particularly if that group occupies a majority position in the decision-making phase of the General Assembly. Discussing the various instruments of influence in the United Nations political system, Kay states that resolutions:

„ . . . are viewed as the logical end toward which much of the activity of the Member States is directed. The activities of the permanent missions, the caucusing groups, and the various organs of the United Nations all bear the imprint of this ultimate ,parliamentary test‘ „<sup>10</sup>.

However, sponsorship of proposals by a delegation can be used for other purposes than goal achievement of the group(s) to which it belongs. Frequency of sponsorship may therefore not be used as an indicator of leadership in a group without further evidence that sponsorship has indeed been used for goal achievement of the group. We solved this problem in two different ways. First, initiative with respect to common group goals is indicated on the basis of the *network of co-sponsorship relations*, because *co-sponsorship* indicates at least compatible, but very often common policy preferences between the co-sponsoring delegations<sup>11</sup>. A central position of a delegation in this network of co-sponsorship relations between the developing nations or groups within that group indicates that such a delegation participated actively in the decision-making together with other group members. The number of proposals co-sponsored by a pair of delegations in that network indicates the extent to which these delegations jointly exercised leadership over the range of group-goal-

<sup>9</sup> One of the research aims of the larger study is a comparison of policy locations and leadership positions of delegations within the group. It required a research design in which selection of roll calls and proposals are closely related to each other. Therefore we selected roll calls concerning only those proposals that were sponsored by delegations and only those proposals that in whole or in part were subjected to a roll call vote. The sponsorship data were coded by Bronneman-Helmerts, H. M., at that time assistant at the Institute for Political Science, and Stokman.

<sup>10</sup> Kay, David A., *The New Nations in the United Nations, 1960–1967*, New York 1970, p. 41.

<sup>11</sup> Mower, A. Glenn, *The Sponsorship of Proposals in the United Nations General Assembly*, in: *Western Political Quarterly*, 15 (1962), p. 662; Kay, Nations, p. 41.

related issues. For the detection of different leadership dimensions that number of co-sponsored proposals can therefore be used as a *measure of similarity* between pairs of delegations. This approach will be followed in the next section. Even if we would find a unidimensional solution, it does not necessarily guarantee the existence of one hierarchy of leadership. A one-dimensional solution might well be reached, because of the existence of different leadership clusters with large intercluster distances but small intracluster distances. If we find such an unidimensional solution, our second approach consists of a specification of a cumulative leadership model. We shall investigate whether there existed a *hierarchy of leadership* within a group. Such a hierarchy implies active participation of certain delegations regarding all group goals, i. e. general leadership. In the case of our sponsorship data we might speak of *one* hierarchy of leadership in a group, if sponsorship is cumulative: leaders in the group will sponsor a large number of proposals over the whole range of group-goal-related issues. „Followers“ in the group are more reluctant to sponsor proposals; they only become involved, if proposals are rather salient to the group goals. The more salient a proposal to the group goals, the more delegations will co-sponsor. If delegations lower in the hierarchy sponsor a proposal, the leaders of the group will also very likely be on the list of sponsors. This approach will be followed in section 3. It should be noted that the two approaches are related to one another. Delegations form a cumulative scale, a hierarchy of leadership, only if they have more co-sponsorship relations with one another in the network of co-sponsorship relations than expected in the case of random sponsorship. The scale analysis can therefore be seen as a special analysis on the network of co-sponsorship relations<sup>12</sup>.

*Group cohesion* among the developing nations is associated with both group functions. Here we only consider group cohesion with respect to the first group function. It is denoted *sponsorship cohesion*. It is reflected in the structure of the network of co-sponsorship relations among developing nations. If that co-sponsorship network contains different, mutually unconnected or loosely connected centers, it reflects a low degree of sponsorship cohesion. We then have different leadership dimensions among the developing nations. Another measure of sponsorship cohesion is contained in the results of cumulative scaling of sponsorship. If all delegations of a group form one cumulative scale, one hierarchy of leadership, it indicates a high degree of sponsorship cohesion; if the group is split over different negatively correlated scales, it indicates low sponsorship cohesion.

<sup>12</sup> Stokman, Roll call. See also Stokman, Frans N., Graph Theoretical Elaboration of Cumulative Scaling Techniques. This article is reproduced in this volume.

## 2. Co-Sponsorship Dimensions

For each period of analysis we can determine the different leadership dimensions among developing nations by a multidimensional analysis of the matrix of co-sponsorship relations among these nations. As we stated above, for each pair of developing nations these matrices contain the number of co-sponsorship relations over the selected proposals in a given period of analysis as a measure of similarity. For each of the periods the leadership dimensions could well be determined by a separate multidimensional analysis of the corresponding matrix. However, to interpret the recurrence of the same dimensions and their importance in the different periods as well as to circumvent interpretation problems in terms of rotation, reflection etc., we shall apply a method of multidimensional scaling, which gives directly comparable leadership dimensions over the different periods: INDSCAL (individual differences scaling).

The model assumes that different individuals perceive a set of  $n$  stimuli in terms of a common set of dimensions, but that these dimensions are differentially important or salient in the perception of different individuals<sup>13</sup>. In our case we assume a common set of leadership dimensions over the four periods of analysis, but these leadership dimensions are differentially important or salient in the different periods of analysis. In this application of the model the stimuli are developing nations and the individuals are periods of analysis. In the common or „group“ space the coordinates of the developing nations are given on the different leadership dimensions; for each of the periods these leadership dimensions are weighted according to saliency of the dimensions; these weighting factors are given in the „subject“ space of the model<sup>14</sup>.

A first impression of the main leadership dimensions among developing nations over the period 1950–68 can be obtained from the INDSCAL solution over all four

<sup>13</sup> Carroll, J. D., and Chang, J.-J., Analysis of Differences in Multidimensional Scaling Via an N-Way Generalization of „Eckart-Young“ Decomposition, in: *Psychometrica*, 35 (1970), pp. 283–319; Carroll, J. Douglas, Individual Differences and Multidimensional Scaling, in: Shepard, Roger N., et al. (eds.), *Multidimensional Scaling. Theory and Application in the Behavioral Sciences*, vol. 1, New York 1972, pp. 105–155.

<sup>14</sup> One may question whether frequency of co-sponsorship is a good measure of similarity that can be used for INDSCAL. Frequency of co-sponsorship between two delegations A and B is dependent of the marginals (the frequency of sponsorship of the two delegations): it can not exceed the frequency of sponsorship of the least active delegation. Some researchers advocate the use of a one-way measure of association (for example Yule's Q) as a measure of similarity to avoid this dependency upon the marginals (Weisberg, Herbert F., *Dimensional Analysis of Legislative Roll Calls*, Univ. Microfilms, Ann Arbor/Mich. 1968; MacRae jr., Duncan, *Issues and Parties in Legislative Voting. Methods of Statistical Analysis*, New York 1970). However, if we had followed this approach, the similarity between two delegations would have been larger to the degree that their sponsorship is more cumulative. This has nothing to do with leadership: sponsorship between two delegations might well be perfectly cumulative, whereas they sponsor only

periods of analysis. Because the model requires the same set of stimuli (developing nations) over all individuals (periods of analysis), this INDSCAL solution is only based on those developing nations, which were a member state of the United Nations during the whole period 1950–68. This caused a drastic reduction of the Afro-Asian nations in this analysis: besides 20 Latin-American countries, only 15 Afro-Asian nations (of which only 3 African nations) could be considered in this analysis. In later analyses the new Afro-Asian member states will also be taken into consideration.

In Table 1 the twodimensional INDSCAL solution over the four periods of analysis is given. It accounts for 50 % of the variance in the data, which was not substantially less than for the three and four dimensional solutions (respectively 54 % and 57 %). On the basis of the weighting factors (Table 1 b) we see that the saliency of the first dimension steadily decreased from .58 to .11, whereas the saliency of the second dimension steadily increased from .22 to .74. On the first dimension of the „group“ space (Table 1 a) we see a number of clusters of developing nations. Most Latin-American delegations have positive coordinates, except for Guatemala, Mexico and Bolivia; most Afro-Asian delegations have strongly negative coordinates, but Yugoslavia, Ethiopia, Liberia and Thailand have positive ones. This dimension is particularly salient in the first period of analysis (1950–55). In that period particularly Guatemala pursued an independent foreign policy until the American intervention in Guatemala in 1954, resulting in a coordinate on the first dimension close to those of the main Afro-Asian countries. The positive coordinates of Yugoslavia and the two more traditional African countries Ethiopia and Liberia in the period 1950–55 might be due to the fact that Afro-Asian group formations was not yet

a limited number of resolutions. In view of our theoretical objectives the best solution would probably have been the Jaccard index. If the cells in the two-by-two tabulation of sponsorship by two delegations are denoted:

	+	–
+	a	b
–	c	d

The Jaccard index is defined as follows:

$$\frac{b + c}{a + b + c}$$

Jaccard's index ignores the cell d. It assumes that that cell does not contain information: similarities of distances can only be based on the number of proposals, sponsored by at least one of the two delegations. If neither delegation sponsored a resolution, that resolution does not contain information about the similarity of distance between the two delegations.

Wolters introduced this measure for the analysis of roll calls in the Dutch parliament (Wolters, Menno, *The Unfolding of Roll-Calls and Thermometer Scores*, Paper presented for the IPSA-Congress at Edingburgh, 1976). In a later version of our paper we intend to perform the INDSCAL analyses on this Jaccard index.

very strong. Only the Arab countries already operated as a caususing group from the beginning of the period. Moreover, Afro-Asian group formation at the end of the period was primarily an Asian enterprise, at that time initiated by the main non-aligned countries Burma, Ceylon, India and Indonesia, which were the main con-

*Table 1: Two Dimensional INDSCAL Solution Over Four Periods of Analysis*

a. coordinates of developing nations in the „group“ space

	dim. 1	dim. 2
Argentina	.0444	.1295
Bolivia	-.0193	.1561
Brazil	.1028	.0850
Chile	.1066	.0672
Colombia	.1320	.1384
Costa Rica	.1352	.1574
Cuba	.0782	.0789
Dom. Rep.	.1159	.1755
Ecuador	.1450	.1296
El Salvador	.0198	.1999
Guatemala	-.1452	.1987
Haiti	.0060	.1448
Honduras	.1587	.1791
Mexico	-.0029	.1189
Nicaragua	.2095	.1545
Panama	.2630	.0869
Paraguay	.2251	.1431
Peru	.2546	.1377
Uruguay	.0382	.1503
Venezuela	.1391	.1086
Yugoslavia	.0828	-.2146
Ethiopia	.1916	-.3010
Liberia	.0716	-.2464
UAR	-.2286	-.2611
Afghanistan	-.2183	-.2227
Burma	-.2486	-.1431
India	-.2016	-.2623
Iran	-.1531	-.1217
Iraq	-.2323	-.2504
Lebanon	-.2469	-.1492
Pakistan	-.2158	-.1890
Philippines	-.2437	-.0488
Saudi Arabia	-.2474	-.1678
Thailand	.0836	.0025
Yemen	-.1998	-.1645

b. weighting factors of the periods of analysis (coordinates in the „subject“ space)

	dim. 1	dim. 2
1950–55	.58	.22
1956–59	.22	.51
1960–63	.17	.64
1965–68	.11	.74

venors of the Bandung Conference in 1955. Yugoslavia and the African countries became involved in Afro-Asian group formation at a later time. The second dimension of the „group“ space gives a perfect split between Latin America and Afro-Asia: all Latin American delegations have positive coordinates on this dimension; all Afro-Asian delegations (except Thailand) have negative coordinates. The coordinates on this second dimension are rather strongly correlated with those on the first dimension ( $r = .60$ ) because the main Latin American and Afro-Asian countries were already splitted on the first dimension. Nevertheless, the increasing importance of this dimension indicates, that over the 1950's and 1960's the split between Latin America and Afro-Asia became more important, whereas separate leadership clusters within the two groups disappeared over time.

The conclusion that different leadership dimensions within Latin America and Afro-Asia diminished in importance over the period 1950–68 was based on an analysis over developing nations which were member states of the United Nations over the whole period. In this analysis we could consider almost all Latin American countries, but only a small number of Afro-Asian delegations. This conclusion should therefore particularly be substantiated for the Afro-Asian group, because of the large number of new Afro-Asian member states, which entered the United Nations after 1950. We therefore performed a second INDSCAL analysis for 44 Afro-Asian delegations over the last two periods of analysis (1960–63 and 1965

*Table 2: Two Dimensional INDSCAL Solution Over the Periods 1960–63 and 1965–68 for 44 Afro-Asian Delegations*

a. coordinates of Afro-Asian nations in the „group“ space

	dim. 1	dim. 2
Cyprus	-.0381	.1114
Yugoslavia	.1280	.2154
Cameroun	.0213	-.2231
Central Afr. Rep.	-.0886	-.2859
Chad	-.0690	-.2252
Congo Brazz.	.0326	-.2650

Congo Dem.	.0778	-.0585
Dahomey	.0178	-.2732
Ethiopia	.0934	.1207
Gabon	-.2386	-.2486
Ghana	.1392	.0425
Guinea	.1555	.0264
Ivory Coast	.0127	-.2609
Liberia	.0860	.0642
Libya	.0939	.0544
Madagascar	-.0718	-.2380
Mali	.1496	.0276
Morocco	.0954	.0660
Niger	.0754	-.2064
Nigeria	.1282	-.0034
Senegal	.0472	-.1399
Somalia	.1092	-.0111
Sudan	.1430	.0448
Togo	.0932	-.1264
Tunesia	.1368	.0552
UAR	.1509	.0748
Upper Volta	.0607	-.1810
Afghanistan	.0585	.1075
Burma	-.2267	.1557
Cambodia	-.3863	.0865
Ceylon	.0091	.1156
India	.1269	.0810
Iran	-.0538	-.0118
Iraq	.1227	.1012
Jordan	-.0206	.1634
Laos	-.4876	.0412
Lebanon	-.0739	.2076
Malaysia	-.2428	.0618
Nepal	-.0080	.1457
Pakistan	.0704	.1005
Philippines	-.1757	-.0349
Saudi Arabia	.0413	.1584
Thailand	-.3305	.1649
Yemen	.0354	.1987

b. weighting factors of the periods of analysis (coordinates in the „subject“ space)

	dim. 1	dim. 2
<hr/>		
1960-63	.32	.61
1965-68	.67	.22
<hr/>		



—68). This analysis provides us the leadership dimensions within the Afro-Asian group after the entrance of a large number of formerly French African member states in 1960. In Table 2 we give the two dimensional INDSCAL solution over the periods 1960–63 and 1965–68. It accounts for 49 % of the variance in the data. On the basis of the weighting factors we see that the first dimension is particularly salient in the period 1965–68, whereas the second dimension is particularly salient in the period 1960–63 (see Table 2 b). The coordinates of the Afro-Asian delegations on the two dimensions in the „group“ space are almost completely uncorrelated ( $r = .01$ ). On the second leadership dimension most formerly French African nations have negative coordinates, particularly the African nations of the so-called *Brazzaville group*.

The Brazzaville group of African States got its name from the Brazzaville Conference in December 1960. This group consisted of twelve states, in essence French Community States<sup>15</sup>. Mauritania, Algeria, and Congo were the main topics on the series of conferences between these states. They favored Mauritanian membership in the United Nations; rejected demands for a referendum in Algeria, but urged France to enter negotiations with the F.L.N.; and supported a round-table conference of the various political factions in the Congo. In the United Nations they constituted an informal group from September 1960 until September 1961<sup>16</sup>. This second leadership dimension is only important in the period 1960–63, because African unity was reestablished in 1963 at the Conference of Addis Abeba, where the Organization of African Unity (OAU) was founded.

It is more difficult to interpret the first leadership dimension, without further analyses. On the basis of the analyses, reported in the next section, we can however, that this dimension is an *activity* dimension: the most actively sponsoring Afro-Asian delegations have high positive coordinates on this dimension, whereas the least active delegations have strongly negative coordinates; this dimension distinguishes the Afro-Asian leaders from the Afro-Asian followers (see particularly Table 6 in Section 3). The exclusive prominence of this dimension in the period 1965–68 indicates that Afro-Asian sponsorship is unidimensional in the period 1965–68. It substantiates our former conclusion that no distinct leadership clusters exist within the Afro-Asian group, even if we take into consideration the large number of new Afro-Asian countries. Leadership among developing nations is indeed increasingly organized along the Latin American/Afro-Asian leadership dimension, after a short period in which controversies among newly independent African nations played an important role. In the next section this will be corroborated by the analysis of cumulative sponsorship dimensions among the developing nations. Moreover, that method of analysis enables us to identify the leaders among the developing nations.

<sup>15</sup> The Brazzaville group consisted of Cameroon, Dahomey, Ivory Coast, Niger, Upper Volta, Madagascar, Mauritania, Gabon, Congo Brazzaville, Chad, Central African Republic and Senegal. Mauritania was not a member state during the whole period 1960–63.

<sup>16</sup> Hovet, Thomas, *Africa and the UNO*, Evenston/Ill. 1963, p. 91.

### 3. Sponsorship Dimensions

A second possibility of detecting leadership dimensions among developing nations consists of specification of a cumulative leadership model. The cumulative leadership model is based on the notion of the existence of one hierarchy of leadership within a group. When do we speak of one leadership hierarchy within a group? We can not speak of one hierarchy if there is a division of labour in the group, in which one subgroup of delegations is active regarding a certain set of group goals, whereas other subgroups are concerned with other sets of group goals. In such a situation we shall probably discover several hierarchies of leadership, depending on the kind of group goals we consider. This can be denoted as issue specific leadership within the group. One hierarchy of leadership within a group implies active participation of certain delegations regarding all group goals, *i. e.* general leadership. Leaders in the group are then active over a broad field of group goal related activities, whereas the „followers“ in the group become only involved, if more central group goals are concerned. In case of our sponsorship data we might speak of one hierarchy of leadership in a group, if sponsorship is cumulative: leaders in the group will sponsor a large number of proposals over the whole range of group related issues. The more central a proposal is related to the group goals, the more delegations will co-sponsor. If delegations, lower in the hierarchy sponsor a proposal, the leaders of the group will also very likely be on the list of sponsors. This model of a hierarchy of leadership is equivalent to a stochastic cumulative response model (a stochastic version of the Guttman model), which was developed by Mokken<sup>17</sup>. The items in the model of Mokken are here delegations. The respondents or subjects in the model of Mokken are now proposals<sup>18</sup>. The trace line in the model of Mokken represents the probability of a positive response by a subject to an item, whereas in our application a trace line represents the probability of sponsorship by a delegation. For our analyses on hierarchies of leadership we can therefore apply the theory and procedures of scale analysis, as developed by Mokken.

We make only some remarks on these theory and procedures, referring the reader to Mokken<sup>19</sup> for more detailed information. For a hierarchy of delegations the fractions of the proposals, sponsored by a delegation, can be used for the ordering of the delegations in the leadership hierarchy: the higher the fraction of sponsored proposals, the higher the position in the hierarchy. This fraction is known as the

<sup>17</sup> Mokken, R. J., *A Theory and Procedure of Scale Analysis*. With applications in political research, The Hague 1970.

<sup>18</sup> In our leadership model we reversed the active and passive set of the scaling model: respondents are here proposals; items are now delegations. In this way we directly get cumulative scales of delegations (hierarchies of leadership). From another theoretical perspective the scaling model might be applied on this sponsorship data to get cumulative scales of proposals over sponsors. Delegations are then considered as respondents and proposals as items.

<sup>19</sup> Theory, in particular chapter 2, 4 and 5.

item difficulty. Here we see, that *frequency of sponsorship can be used to indicate leadership in a group, if the data fit the cumulative leadership model*. The proposals can be given a score equal to the number of sponsoring delegations. The order of this score gives an ordering of the proposals relative to centrality to group goals.

For a hierarchy of  $k$  delegations, it can be proved that for all delegations  $i, j$  ( $i, j = 1, 2, \dots k$ ) the *probability of co-sponsorship is higher than the probability expected on the basis of random sponsorship*: sponsorship of the different delegations in a hierarchy of leadership is positively correlated. This implication of a positive correlation was used by Mokken to define a criterion of scalability. Mokken used the coefficient  $\Phi/\Phi_{\max}$  as a measure of positive correlation between any pair of items (delegations). For a scale as a whole, Loewinger's coefficient of homogeneity,  $H$ , is used as a criterion of scalability in the sense of monotone homogeneity.  $H$  can be written as a weighted function of the  $\Phi/\Phi_{\max}$ , the  $H_{ij}$ 's, for all item pairs. Mokken also introduced an item coefficient  $H_i$ , in terms of which the scalability of each item (delegation) relative to the other items (delegations) can be evaluated.  $H_i$  can also be written as a weighted function of the  $H_{ij}$ -coefficients ( $\Phi/\Phi_{\max}$ ) of that item with all other items<sup>20</sup>. A scale, in our case a hierarchy of leadership, is then defined as follows: „a scale is a set of items which are all positively correlated and with the property that every item coefficient of scalability ( $H_i$ ) is larger than or equal to a given positive constant ( $c$ )“<sup>21</sup>. If all  $H_i \geq c$ . Mokken<sup>22</sup> suggested the following degrees of scalability

- a.  $.50 \leq H$ : a strong scale;
- b.  $.40 \leq H < .50$ : a medium scale;
- c.  $.30 \leq H < .40$ : a weak scale.

Thus, the coefficients  $H$  and  $H_i$  indicate the degree of hierarchization and the extent to which policy goals of the different nations overlap, whereas the frequency of sponsorship of a delegation indicates leadership in a group, if the data fit the cumulative leadership model.

On the basis of this theory of scale analysis Mokken proposed a class of scaling procedures<sup>23</sup>. One of these procedures consists of *multiple scaling*, the construction of a number of scales from a given pool of items. The multiple scaling procedure selects items (delegations) from a set of items (delegations) in such a way that the scale coefficient  $H$  is maximized. If no other items can be added to a scale on the basis of the chosen threshold level of the  $H$  and  $H_i$  coefficient, a second scale is constructed from the remaining items in the same way. This process is repeated until no other scales can be found which satisfy the threshold level<sup>24</sup>.

<sup>20</sup> Op. cit., pp. 150 passim.

<sup>21</sup> Op. cit., p. 184.

<sup>22</sup> Op. cit., p. 185.

<sup>23</sup> Op. cit., pp. 187–199.

<sup>24</sup> For each period the analysis is performed over all states, which were member state of the United Nations during the whole period. However, delegations, which sponsored less than two proposals in a period of analysis, were eliminated for that period. Their correlation coefficients

Table 3: Cumulative Sponsorship Scales of Delegations; Period 1950-55

	fraction of sponsored proposals	$H_i$
<i>a. Afro-Asian scale</i>		
Iran	.04	.89
Yemen	.07	.79
Saudi Arabia	.10	.67
Burma	.11	.59
Pakistan	.12	.56
Indonesia	.15	.62
Syria	.16	.54
UAR	.17	.57
Iraq	.18	.54
coefficient of scalability for the whole scale $H = .61$		
<i>b. First Latin American scale</i>		
Nicaragua	.01	1.00
Paraguay	.02	.89
Panama	.02	.87
Honduras	.02	.66
Costa Rica	.04	.62
Peru	.05	.64
Ecuador	.09	.62
Brazil	.14	.65
coefficient of scalability for the whole scale $H = .70$		
<i>c. Second Latin American scale</i>		
Bolivia	.05	.57
Mexico	.07	.50
Guatemala	.10	.54
coefficient of scalability for the whole scale $H = .54$		

In Table 3 the leadership hierarchies are given for the *period 1950-55*. The three scales resulted from multiple scaling with a lowerbound of .50 for the item coefficients  $H_i$ . Therefore, the three scales in Table 3 are strong scales. The Arab delegations took an important position in the *Afro-Asian hierarchy* of 9 delegations ( $H = .61$ ). Iraq, the UAR and Syria were the three main leaders of the hierarchy,

$H_{ij}$  can only be negative or one. By consequence, they were quite often selected as one of the first items in the search procedure. This disturbed the results of the multiple scaling procedure. Only scales of developing nations are reported.

sponsoring respectively 18, 17 and 16 percent of all proposals. Burma and Indonesia violated the requirement of double monotony: their trace lines intersect those of the other delegations<sup>25</sup>. It confirms that the hierarchy is in essence an Arab leadership hierarchy instead of an Afro-Asian one. If delegations with item coefficients between .50 and .30 were allowed, the 9 item scale became a scale of 15 delegations ( $H = .51$ ). All 15 delegations were Afro-Asian delegations. The leader in this larger scale of 15 Afro-Asian delegations is now India, which sponsored 20 percent of all proposals. We reported the 9 item strong scale in Table 3 instead of the hierarchy of 15 delegations, because there were many disturbances of the double monotony in the larger scale. Only two Afro-Asian delegations were not contained in the larger scale: Ethiopia was eliminated because it sponsored only one proposal in this period; Yugoslavia did not scale with the other Afro-Asian delegations; Yugoslavia sponsored 9 percent of all proposals.

The results of the scale analysis confirms the low level of group formation among the Afro-Asian delegations, which we already noticed in the previous section. The Arab countries already operated as a permanent caucusing group from the beginning of the period. Only in the second half of the period India really managed to organize an Afro-Asian group within and outside the United Nations (Bandung Conference in 1955), of which it became one of the main leaders. Indeed, the hierarchization of the Arab group, but not of the overall Afro-Asian group, was strong enough to form a strong cumulative sponsorship scale. In terms of sponsorship India was the most active Afro-Asian delegation (it sponsored 20 percent of the proposals), but at least during the first years of the period 1950–55 it was an active Afro-Asian delegation without an effectively operating group of delegations around it.

In the period 1950–55 two *Latin American hierarchies* of leadership existed. The first, dominant Latin American hierarchy ( $H = .70$ ) consisted of 8 delegations. Brazil is the leader; it sponsored 14 percent of all selected proposals. The hierarchy can be extended to 11 delegations ( $H = .51$ ), if also delegations with item coefficients between .50 and .30 are allowed. Venezuela, Uruguay and Cuba are then added to the scale; they sponsored respectively 4, 6 and 11 percent of the proposals. They disturb the requirement of double monotony and are therefore not acceptable. The second Latin-American scale consisted of Bolivia, Mexico and Guatemala ( $H = .54$ ). At a lowerbound of .30 El Salvador and Yugoslavia were added. This resulted in a weak 5-item scale ( $H = .38$ ). Sponsorship of delegations in the dominant Latin American scale is completely independent of that in the Afro-Asian scale: the item

<sup>25</sup> The cumulative scaling model does not only require that the trace lines increase monotonely along the underlying continuum, but also that trace lines do not intersect, i. e. that the requirement of double monotony or holomorphism is fulfilled. It does not specify the function of the trace lines, however. If a set of items forms a scale following the criteria of scalability (a set of delegations forms an hierarchy of leadership following these criteria), it is still possible that the different trace lines intersect. This can be checked as described by Mokken (op. cit., pp. 180–182).

coefficients  $H_1$  of the delegations in the dominant Latin American scale with respect to the Afro-Asian scale range from  $-.09$  to  $.15$ . It clearly shows that the group goals of this dominant Latin-American group do not coincide with those of the Afro-Asian delegations. This is not the case for the second Latin American group. Sponsorship of these Latin American delegations is not quite independent of that of the Afro-Asian delegations: 6 delegations of the Afro-Asian scale have item coefficients of  $.30$  or higher with respect to this second Latin American hierarchy. This second Latin American hierarchy is closer related to the Afro-Asian scale than to the dominant Latin American scale: only one delegation of the dominant Latin American scale had an item coefficient of  $.30$  or higher with the second Latin American scale.

We conclude, that there existed three main leadership hierarchies among the developing nations in the period 1950–55: one predominantly Arab group and two Latin American groups. India was very active among the Afro-Asian delegations, but in the first years of the period it had not yet organized the Afro-Asian group effectively, Brazil and Ecuador were the leaders of the dominant Latin American group. Sponsorship of this group was completely independent of that of the Afro-Asian group. This group had therefore its own group goals, different from those of the Afro-Asian group. Guatemala and Mexico are the two main delegations in the second Latin American group. Until the American intervention in Guatemala in 1954, these two countries pursued an independent foreign policy. These results correspond remarkably well with the coordinates of the developing nations on the first leadership dimension in the two dimensional INDSCAL solution over the four periods of analysis (see Table 1). As we stated above, that dimension was particularly salient in the period 1950–55.

To limit the size of this paper we do not extensively report the cumulative scales of delegations in the *period 1956–59*, referring the reader to Stokman<sup>26</sup>. For a comparison with the INDSCAL analyses we rather prefer to report the cumulative scaling results of the periods 1960–63 and 1965–68, because of the Brazzaville cluster in the period 1960–63 and the prominence of the Latin American/Afro-Asian dimension in the period 1965–68. Multiple scaling of sponsorship in the period 1956–59 showed the existence of a real strong Afro-Asian group under the leadership of Ceylon, India, Indonesia and Burma, the main conveners of the Bandung Conference in 1955. The Western aligned Afro-Asian delegations however formed a separate group or did not belong to any of the Afro-Asian groups. It indicates that in this period problems of nonalignment versus alignment in Cold War had consequences for leadership positions and structures among developing nations with respect to colonial and socio-economic issues. Group formation of Latin American delegations was very weak with respect to these issues. As far as Latin American groups existed, they took more pro-western than Afro-Asian policy positions.

<sup>26</sup> Stokman, Roll Calls.

In the period 1960–63 two main leadership hierarchies were found: an Afro-Asian scale of 44 delegations ( $H = .64$ ) and a Latin American scale of 8 delegations ( $H = .60$ ). Both scales are given in Tabel 4. They resulted from multiple scaling with lowerbound level of  $H_i = .50$ . Only three Afro-Asian delegations did not belong to the Afro-Asian hierarchy: the Philippines, Thailand and Yugoslavia. The Philippines had an item coefficient  $H_i = .497$  with respect to this scale; in fact, it can be considered as a member of the hierarchy; it sponsored 15 percent of the proposals. Thailand and Yugoslavia had negative correlations with some other Afro-Asian delegations. However, with the leaders of the Afro-Asian hierarchy both delegations had  $H_{ij}$  coefficients above .50. Yugoslavia and Thailand did not belong to the Afro-Asian hierarchy, because they constituted different wings in this period. This can be seen from the following multiple scaling results. Apart from the two hierarchies reported in tabel 4, multiple scaling resulted in two other hierarchies. The first one (with Thailand) consisted of Colombia, Chile, Thailand and the Philippines ( $H = .61$ ); the second one (with Yugoslavia) consisted of Czechoslovakia, Haiti, Cuba and Yugoslavia ( $H = .75$ ). We did not report these two scales as separate hierarchies in Tabel 4, because all leaders from Indonesia to Ghana (with only one exception) and some other delegations had item coefficients above .50 with respect to both scales. They can therefore not be considered as separate hierarchies. These results strongly indicate *general* leadership in the Afro-Asian group with different wings among the followers. The four most active delegations in the Afro-Asian hierarchy were four members of the *Casablanca group*. The Casablanca group of African states got its name from the Casablanca Conference, held in January 1961. These African states<sup>27</sup> held more radical views on African issues and nonalignment than the Brazzaville group. In the Congo crisis they all favored the Lumumba faction and demanded UN support for that faction. They declared their intention to withdraw their troops from the ONUC. With respect to Mauritania they supported the claim of Morocco, which considered Mauritania as a part of Great-Morocco. In the United Nations about twelve formal meetings of the group were held between January 1961 and November 1962<sup>28</sup>.

In the selection procedure the former French African colonies were added to the scale as the 31st to 43rd delegations in the scale. Only one other Afro-Asian delegation was added to the scale in a later step: Iran was added as the last delegation in the hierarchy. The item coefficient of Upper Volta, which was added as the 31st delegation, was .59 at the moment of selection, but increased to .65 after selection of Gabon as 43rd delegation. Also some other item coefficients of former French African colonies raised during the selection procedure. Most former French African colonies systematically violated the requirement of double monotony in the final hierarchy of 44 delegations. This is in particular the case for the delegations, which constituted the Brazzaville group. The increasing item coefficients and the disturbance

<sup>27</sup> Guinea, Mali, Algeria, Morocco and the UAR. Algeria was not a member state of the United Nations during the whole period 1960–63.

<sup>28</sup> Hovet, Africa, p. 98.

Table 4: Cumulative Sponsorship Scales of Delegations; Period 1960-63

	fraction of sponsored proposals	$H_i$		fraction of sponsored proposals	$H_i$
<i>a. Afro-Asian scale</i>					
Laos	.03	.79	Pakistan	.19	.55
Turkey	.03	.48	Niger	.19	.66
Japan	.03	.60	Togo	.21	.60
Cyprus	.08	.59	Somalia	.22	.69
Yemen	.12	.54	Nepal	.22	.57
Gabon	.12	.54	Ceylon	.24	.66
Central African Rep.	.12	.61	Burma	.24	.60
Madagascar	.12	.59	Senegal	.24	.65
Saudi Arabia	.13	.58	Ethiopia	.24	.62
Lebanon	.15	.55	Libya	.26	.69
Dahomey	.15	.60	Liberia	.26	.60
Congo-Brazzaville	.15	.62	Tunesia	.28	.59
Iran	.15	.52	Indonesia	.29	.69
Jordan	.16	.61	Sudan	.29	.66
Malaysia	.16	.64	Mali	.29	.73
Congo-Dem. Rep.	.16	.60	India	.31	.69
Ivory Coast	.16	.60	Iraq	.31	.69
Cameroon	.17	.60	Nigeria	.31	.72
Cambodia	.17	.63	Morocco	.32	.77
Chad	.17	.58	Guinea	.34	.76
Upper Volta	.18	.65	UAR	.35	.78
Afghanistan	.19	.58	Ghana	.38	.81
coefficient of scalability for the whole scale $H = .64$					
<i>b. Latin American scale</i>					
Costa Rica	.02	.60			
Ecuador	.02	.55			
Guatemala	.02	.70			
El Salvador	.03	.54			
Peru	.04	.52			
Uruguay	.04	.68			
Argentina	.07	.56			
Brazil	.07	.62			
coefficient of scalability for the whole scale $H = .60$ .					



Table 5: Cumulative Sponsorship Scales of Delegation; Period 1956-68

	fraction of sponsored proposals	$H_i$		fraction of sponsored proposals	$H_i$
<i>a. Afro-Asian scale</i>					
Laos	.03	.79	Liberia	.34	.52
Turkey	.05	.65	Niger	.34	.61
Singapore	.08	.59	Burundi	.35	.62
Gambia	.08	.68	Uganda	.35	.66
Burma	.09	.59	Ethiopia	.36	.65
Malaysia	.10	.65	Morocco	.36	.64
Thailand	.11	.54	Tunesia	.37	.65
Gabon	.12	.59	Somalia	.37	.60
Trinidad & Tobago	.14	.54	Libya	.37	.63
Mongolia	.16	.65	Congo-Dem. Rep.	.38	.65
Central African Rep.	.19	.69	Togo	.39	.63
Chad	.20	.67	Sierra Leone	.41	.65
Lebanon	.21	.58	Pakistan	.42	.62
Madagascar	.24	.59	Sudan	.42	.68
Iran	.24	.50	Kenya	.43	.73
Kuwait	.25	.54	Mali	.44	.67
Jordan	.25	.59	Zambia	.44	.66
Nepal	.25	.58	Mauritania	.44	.63
Cyprus	.28	.65	Iraq	.45	.63
Ivory Coast	.28	.59	Nigeria	.45	.74
Ceylon	.29	.51	Yugoslavia	.45	.60
Congo-Brazzaville	.31	.59	Ghana	.46	.71
Senegal	.31	.60	India	.47	.71
Dahomey	.31	.65	Algeria	.49	.74
Upper Volta	.31	.54	UAR	.50	.71
Rwanda	.32	.59	Syria	.51	.74
Cameroon	.32	.60	Tanzania	.52	.77
Afghanistan	.34	.58	Guinea	.53	.79
coefficient of scalability for the whole scale $H = .63$					
<i>b. Latin American scale</i>					
Paraguay	.04	.91	Peru	.08	.76
Brazil	.05	.61	Costa Rica	.09	.64
El Salvador	.05	.89	Panama	.09	.62
Honduras	.06	.87	Venezuela	.10	.72
Mexico	.06	.57	Uruguay	.10	.78
Bolivia	.08	.75	Argentina	.11	.74
Dominican Rep.	.08	.67	Chile	.11	.60
Guatemala	.08	.82	Colombia	.11	.78

Haiti	.08	.74	Ecuador	.12	.79
Nicaragua	.08	.79			

coefficient of scalability for the whole scale  $H = .74$

---

in the double monotony can be attributed to the extremely high scalability among the Brazzaville delegations: the 11 Brazzaville delegations constituted a very strong hierarchy with a coefficient of scalability  $H = .83$ . They constituted therefore a group within the larger Afro-Asian group.

The second hierarchy in Table 4 is a strong scale of 8 Latin American delegations ( $H = .60$ ). The two largest Latin American countries are the leaders of this scale: Brazil and Argentina. They were not very active, however: they sponsored only 7 percent of all proposals in this period, primarily regarding socio-economic issues. The neutralist policy of President Goulart of Brazil from 1961 to 1964 did therefore not result in a separate hierarchy of leadership, close to the Afro-Asian, as was the case with Guatemala in the period 1950–55; it only resulted in a leadership position of Brazil in the dominant Latin American group, which it lost again in the period 1965 to 1968, as we shall see. One may therefore question whether the foreign policy of President Goulart was so neutralist as it pretended to be. At the lowerbound level of .30 only one other Latin American delegation could be added to this Latin American scale, Haiti. None of the delegations in the Afro-Asian scale had an item coefficient above .26 with respect to this Latin American scale. Sponsorship of Latin American delegations is therefore not related to primarily Afro-Asian proposals in this period. It holds at most for Colombia, Chile, Haiti and Cuba, which were part of the third and fourth scales. The small number of proposals, which were sponsored by these delegations, ranging from 2 to 5 percent of all proposals, make these conclusions tentative, however.

We conclude, that in the period 1960–63 the Afro-Asian delegations formed one main hierarchy under the leadership of the Casablanca countries. Among the Afro-Asian followers we can nevertheless distinguish three different wings: one consisting of Thailand and the Philippines; one consisting of Yugoslavia, and one consisting of the Brazzaville group. Among the Latin American countries there existed one not very active hierarchy of 8 delegations under the leadership of Argentina and Brazil, which was mainly concerned with socio-economic issues. This Latin American group did not fit in the Afro-Asian hierarchy; it had its own group goals regarding colonial and socio-economics issues. This was not the case for four other Latin American countries: Colombia and Chile belonged to the Afro-Asian wing consisting of Thailand and the Philippines; Haiti and Cuba belonged to the Yugoslavia-wing in the Afro-Asian group.

The scale analysis in the *period 1965–68* also gave very remarkable results. The first hierarchy is an Afro-Asian scale, which consisted of 56 delegations. The second hierarchy is a Latin American scale of 19 delegations. Both hierarchies are strong scales (see Table 5). The Afro-Asian hierarchy encompassed 53 delegations of the

Afro-Asian caucusing group. The three other delegations are Turkey, Mongolia and Trinidad and Tobago. Only 6 of the 59 Afro-Asian delegations were not contained in this large Afro-Asian scale. Saudi Arabia ( $H = .49$ ) and the Philippines ( $H_1 = .40$ ) were excluded. The four other Afro-Asian delegations had negative correlations with some other Afro-Asian delegations in the scale. For this reason they were rejected. These four delegations were Cambodia, Malawi, Maldives Islands and Yemen.

One of the most remarkable differences with the preceding period is the scalability of Yugoslavia. For the first time Yugoslavia is contained in the Afro-Asian scale. It sponsored 45 percent of the proposals, which is considerably more than in the preceding periods. Another striking difference between the third and fourth period is the scalability of the Brazzaville group. In the period 1960--63 these delegations disturbed the double monotony because they formed a subgroup within the larger Afro-Asian hierarchy with a very high degree of hierarchization ( $H = .83$ ). In the period 1965--68 the coefficient of scalability of the Brazzaville group is not higher than that of the whole Afro-Asian scale, namely  $H = .65$ .

The delegations of the Casablanca group still belonged to the leaders of the Afro-Asian group, but the Afro-Asian center was extended with a number of other delegations, in particular Tanzania and Syria. For the first time we found one Latin American hierarchy, which encompassed nearly the whole Latin American group. Only three Latin American delegations were not contained in this scale: Cuba, Jamaica and Trinidad and Tobago. Jamaica and Trinidad and Tobago could be added to the scale, if item coefficients between .50 and .30 were allowed. Mexico and to a lesser degree Chile systematically disturbed the double monotony in the Latin American scale. Sponsorship of Latin American delegations was quite independent of that of Afro-Asian delegations. In this period sponsorship of Latin America and Afro-Asia were even negatively correlated: most Afro-Asian leaders had strongly negative item coefficients with respect to the Latin American scale.

We conclude, that in the period 1965--68 two main leadership hierarchies encompassed nearly all developing nations: an Afro-Asian hierarchy encompassed almost all Afro-Asian delegations, a Latin American one almost all Latin American delegations. It was for the first time that the Latin American delegations operated as one group with respect to colonial and socio-economic issues. It was also for the first time, that sponsorship of Latin American delegations was negatively correlated with that of the Afro-Asian delegations. These developments might well be the result of the same developments in Latin America, which led to the „Group of 77“. The common economic problems of the Latin American countries in the 1960's reinforced a group process among the Latin American countries with respect to these socio-economic problems resulting in a more encompassing group and at the same time more clearly defined Latin American group goals. The similarity of the economic problems between Latin America and Afro-Asia in many respects led to the formation of the „Group of 77“ as a coalition of these two groups. This further stimulated group processes within Latin America in this field: only as a distinct, well organized subgroup could the Latin American countries be sure to get the necessary priorities for their own policy preferences within the „Group of 77“. The strongly

regional character of the „Group of 77“ also presumed a distinct Latin American group. The fact that the Latin American hierarchy was particularly active with respect to socio-economic issues can be seen as a further corroboration of this interpretation of Latin American group processes at that time. In each case the results of the scale analyses clearly demonstrate that *the „Group of 77“ must be seen as a coalition between two separate groups: the Latin American group and the Afro-Asian group.*

#### 4. Conclusions

We conclude, that over the 1950's and 1960's *co-sponsorship* among developing nations was structured along one underlying dimension. Over time this underlying dimension took on increasingly the character of a split between Latin American on the one hand and Afro-Asia on the other: different leadership clusters within these two important subgroups gradually disappeared. In Latin America this was the case with Guatemala and Mexico, which followed a policy close to that of the main Afro-Asian delegations in the period 1950–55; in Afro-Asia it holds in particular for Yugoslavia and the Brazzaville group, which no longer existed as separate Afro-Asian leadership clusters in the period 1965–68.

Analysis of *sponsorship* showed that in each period the underlying co-sponsorship dimension consisted of different leadership hierarchies. These leadership hierarchies coincided with the different clusters of delegations on the underlying co-sponsorship dimension. It enabled us to identify the leaders in the different leadership hierarchies and it substantiated our conclusion that leadership among developing nations was increasingly organized into two subgroups: Latin America and Afro-Asia.

The separate Afro-Asian and Latin American leadership hierarchies in the period 1965–68 reflect the importance of the regional groups in the decision-making in the „Group of 77“. Gosovic distinguished three regional groups: Latin America, Africa and Asia. In UNCTAD's main bodies (Conference and Board) even the smallest proposals were referred to these regional groups for clearance<sup>29</sup>. Our results suggest, however, that the relations between the Asian and the African regional groups were quite different from those between Latin America on the one hand and the African and Asian groups on the other hand. At least in the General Assembly, the African and the Asian groups operated as one group, whereas the Latin American delegations still behaved as a separate group within the „Group of 77“, as far as their behavior in sponsorship is concerned.

<sup>29</sup> Gosovic, Branislav, UNCTAD. Conflict and Compromise. The third world's quest for an equitable world economic order through the United Nations, Leyden 1972, p. 206–7.

## Graph Theoretical Elaboration of Cumulative Scaling Techniques

In this paper a graph theoretical elaboration of the *stochastic* cumulative scaling model of Mokken<sup>1</sup> will be given to determine:

- a. cumulative scales of *vertices* on the basis of their relations with other vertices in a simple graph;
- b. cumulative scales of relations in a multigraph.

In Stokman both graph theoretical elaborations are given and applied to (co) sponsorship of resolutions in the United Nations General Assembly to determine leadership structures among developing nations<sup>2</sup>. Felling elaborated the applicability of the *deterministic* cumulative scaling model of Guttman to determine cumulative scales of relations in multigraphs<sup>3</sup>. Graph theoretical elaboration of the stochastic scaling model gives however certain new insights, that were not considered by Felling.

In Section 1 we introduce a number of graph theoretical concepts that will be used in the remainder of the paper. In Section 2 cumulative scales of vertices in a simple graph are treated, in Section 3 cumulative scales of relations in a multigraph. After determination of the different cumulative dimensions of relations in a multigraph, for each dimension a new graph can be generated and analyzed. In Section 4 it will be shown how this can be done.

### 1. Graph Theoretical Concepts<sup>4</sup>

A *graph* is an object, that contains vertices and edges, each edge being incident with one or two vertices<sup>5</sup>. Let us consider, as an example, the graph that consists of the vertices *u, v, w, x, y, z* and the edges *a, b, c, d, e, f, g, h*. The incidence relations are given in Table 1. The graph is given in Figure 1.

<sup>1</sup> Mokken, R. J., A Theory and Procedure of Scale Analysis. With applications in political research, The Hague 1970.

<sup>2</sup> See Stokman, Frans N., Roll Calls and Sponsorship. A methodological analysis of Third World group formation in the United Nations, Leiden 1977.

<sup>3</sup> Felling, A. J. A., Sociaal-netwerkanalyse, Alphen aan den Rijn 1974.

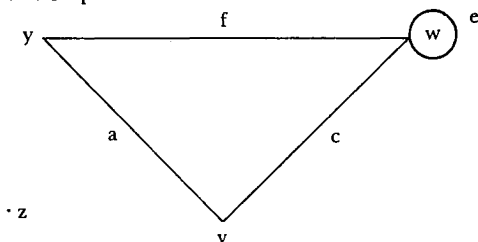
<sup>4</sup> The presentation of the graph theoretical concepts is strongly based on Helmers, H. M., et.al.

Table 1: Incidence Relations

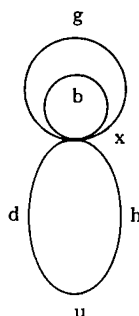
edge	incident with
a	v and y
b	x
c	w and v
d	u and x
e	w
f	y and w
g	x
h	u and x

Source: Helmers et. al., 1975, p. 112.

Figure 1: A Graph



Source: Helmers et. al., 1975, p. 115.



The edges e, g and b are only incident with one vertex. Such edges are called *loops*. The other edges are incident with two vertices. These two vertices are directly connected by such edges. Two directly connected vertices are *adjacent* or *neighbors*. The vertex z has no neighbors. Such a vertex is an *isolated* vertex. The edges d and h are both incident with u and x; such edges are *parallel*. The number of parallel edges between two vertices is the *multiplicity* of that direct connection between two vertices, also called the multiplicity of the edges.

Graven naar macht. Op zoek naar de kern van de Nederlandse economie, Amsterdam 1975, and Stokman, Roll Calls.

For a more elaborate introduction the reader is referred to Öre, Oystein, Graphs and Their Uses, New York 1963; Harary, Frank, Graph Theory, Reading/Mass. 1969, and Harary, Frank, et.al., Structural Models: An Introduction to the Theory of Directed Graphs, New York 1965.

<sup>5</sup> A directed graph is a graph in which each edge has a direction, from one vertex to another. An edge, together with its direction, is called an arc. In this paper we consider only undirected graphs. The stochastic scale model can also be applied in directed graphs; in that case the direction of the arcs should be taken into account.

A *simple graph* is a graph without loops and parallel edges. If parallel edges (can) occur in a graph, such a graph is called a *multigraph*. Information can be added to the edges and/or vertices of a graph. A graph, together with the information associated with its elements, is a *network*. Multigraphs can be represented as networks, if adjacent vertices are connected by only one edge and the multiplicity of the relation is added as information to that edge. In case of different kinds of relations between elements a multigraph can be used as representation by adding the kind of the relation as information to the edges. This situation will be considered in section 3. A graph is a *bipartite graph*, if the set of vertices of the graph is divided in two disjoint, non-empty subsets in such a way that the vertices have no neighbors within their own subset.

A graph or network can be used to *generate* other graphs or networks. Selection, aggregation and induction are the main processes used to generate a new graph or network from an existing one<sup>6</sup>.

A new graph is generated by *selection*, if its elements (vertices, edges) are chosen from the set of elements of the original graph. The new graph then consists of the vertices and edges that satisfy the selection criteria. The selected edges must be edges of the new graph, the vertices incident with each selected edge must be selected vertices. A *subgraph* consists of a subset of the vertices of the original graph and all edges of the original graph of which both incident vertices belong to the selected subset of vertices. A subgraph is generated by deleting a number of vertices and all edges that are incident with these vertices. A *partial graph* consists of all vertices of the original graph and a subset of edges of that graph. A *partial subgraph* is a partial graph of a subgraph.

A new graph is generated by *aggregation of vertices*, if a subset of vertices in the original graph is condensed to a new vertex. Edges with both incident vertices within the subset became loops on the new vertex. Each edge between a vertex within and a vertex outside the subset becomes an edge between the new vertex and the vertex outside the subset. A new graph can also be generated by *aggregation of edges*, for example by condensing parallel edges to one new edge.

*Induction* is the third main process used to generate new graphs. In the case of induction two vertices in the new graph are directly connected by an edge, if these two vertices have a common neighbor in the original graph. The vertices of the induced graph (called inductees) are a subset of the vertices of the original graph. The induction can be limited to common neighbors in a certain subset of vertices of the original graph, called the subset of inductors. Each inductor induces edges between the inductees to which it is adjacent.

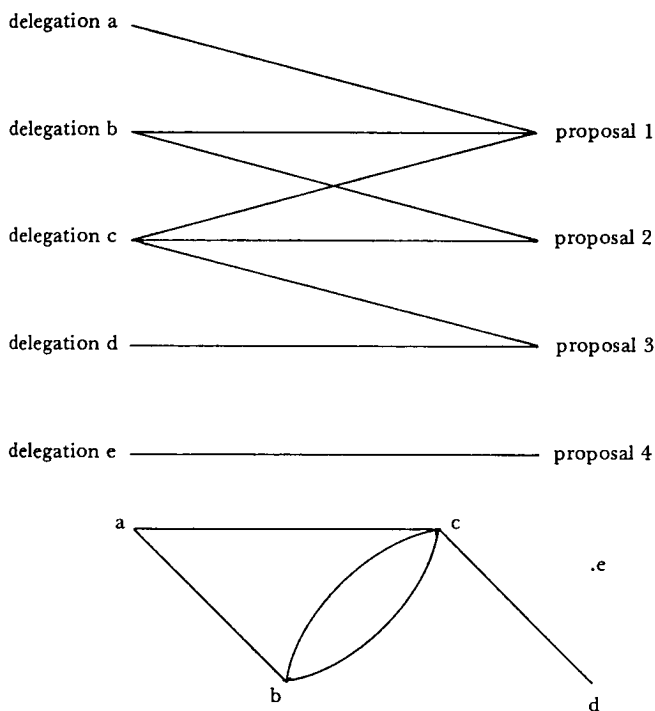
As an illustration, Figure 2 contains a bipartite graph. Each vertex in the first subset represents a delegation, each vertex in the other subset represents a proposal. A delegation and a proposal are connected by an edge if that delegation sponsored that proposal. This graph can be used to generate other graphs or networks, in particular those of co-sponsorship relations between delegations. The subset of in-

<sup>6</sup> Anthonisse, Jac M., and Lageweg, B. J., Graphlib O, Amsterdam 1975.

ductees are then the delegations, the subset of inductors the proposals. By induction a new graph is generated, consisting of relations between delegations. Each proposal induces an edge between each pair of delegations, by which it was sponsored. In the new graph two delegations are directly connected by as many edges as they co-sponsored proposals. The relations in the new graph are therefore the co-sponsorship relations between delegations. In this new graph parallel edges can be aggregated; the multiplicity of the edges (the number of co-sponsored proposals) is associated as information with each edge<sup>7</sup>.

For the *analysis* of a graph or network we consider here only the density, the connectivity and the component density of a graph. We ignore the multiplicity of the edges. If the multiplicity is associated as information with the edges, we therefore ignore that information for the moment. Later in this section we shall take that information into account.

Figure 2: Induction



<sup>7</sup> The bipartite graph between delegations and resolutions could have been represented as a directed graph by associating a direction with each edge, e. g. from delegation to resolution. Analytically this has no meaning because only symmetric relations can be distinguished between delegations after induction, namely co-sponsorship relations.



The *density* ( $a$ ) of a graph is the fraction of adjacent vertices. In a graph with  $p$  vertices and  $b$  neighbor relations (edges), the density is:

$$a = \frac{b}{1/2 p(p-1)} \quad (1)$$

In bipartite graphs vertices of the same subset cannot be adjacent. This can be taken into account by defining the density of a bipartite graph as:

$$a = \frac{b}{q \cdot r} \quad (2)$$

in which  $q$  and  $r$  are the number of vertices in the two subsets.

If two vertices in a graph are not adjacent, they can be connected indirectly, *e. g.* because they have a common neighbor. If two vertices are adjacent, indirect connections between the two vertices can exist as well.

A *path* between vertex  $x$  and vertex  $y$  of a graph consists of an alternative sequence of edges  $l_i$  and vertices  $z_i$ :

$$x, l_1, z_1, l_2, z_2, \dots, z_{k-1}, l_k, y \quad (3)$$

in which  $l_1$  is incident with  $x$  and  $z_1$ ,  $l_2$  incident with  $z_1$  and  $z_2$  etc. and  $l_k$  incident with  $z_{k-1}$  and  $y$ .

The vertices  $x$  and  $y$  are joined by a path through the vertices  $z_i$  and edges  $l_i$ .

A graph consists of one or more *components*. Two vertices belong to the same component if they are joined by a path; if two vertices are not joined by a path, they belong to different components. A graph is *connected* if it has only one component, *i. e.* if every pair of vertices is joined by a path. A component is therefore a maximal connected subgraph of the graph.

The connectivity of a graph is the number of pairs of vertices, joined by a path, as a fraction of all pairs of vertices in a graph. If the graph consists of  $s$  components and the  $i$ -th component consists of  $p_i$  vertices, the connectivity of the graph is:

$$c = \frac{2}{p(p-1)} \sum_{i=1}^s 1/2 p_i (p_i - 1) \quad (4)$$

The connectivity between two disjoint subsets  $Q$  and  $R$  of vertices of a graph is based on all pairs of vertices, of which one vertex belongs to one subset and the other to the other subset. If the  $i$ -th component contains  $q_i$  vertices from the first subset and  $r_i$  vertices from the other subset, the connectivity between the subsets  $Q$  and  $R$  is:

$$c_{QR} = \frac{1}{q \cdot r} \sum_{i=1}^s q_i r_i \quad (5)$$

The density is based on all pairs of vertices that are directly connected; the connectivity is based on all pairs of vertices that are connected, either directly or indirectly.

Therefore we always have:

$$c \geq a$$

with equality if all pairs of vertices within all components are adjacent.

In the density we expressed the number of adjacent vertices as a fraction of all pairs of vertices. We can also express the number of adjacent vertices in a graph as a fraction of all pairs of vertices that belong to the same component, because vertices in different components cannot be adjacent. This *component density* is defined by:

$$a_c = \frac{b}{\sum_{i=1}^s 1/2 p_i (p_i - 1)} = \frac{2b}{\sum_{i=1}^s p_i (p_i - 1)} \quad (6)$$

The relation between the connectivity, the component density and the density is:

$$a = a_c \cdot c$$

For bipartite graphs the component density is defined by:

$$a_c = \frac{b}{\sum_{i=1}^s q_i r_i} \quad (7)$$

Again we have:

$$a' = a_c \cdot cQR$$

One of the possibilities to take the multiplicity of the edges into account in the analyses is a *repeated (stepwise) analysis of the graph for different levels of multiplicity*. From the network we select a partial graph, containing only edges of a certain level of multiplicity or higher, and we analyze that partial graph; we repeat that for different levels of multiplicity. For example, we might analyze the network of co-sponsorship relations for the following levels of multiplicity (m):

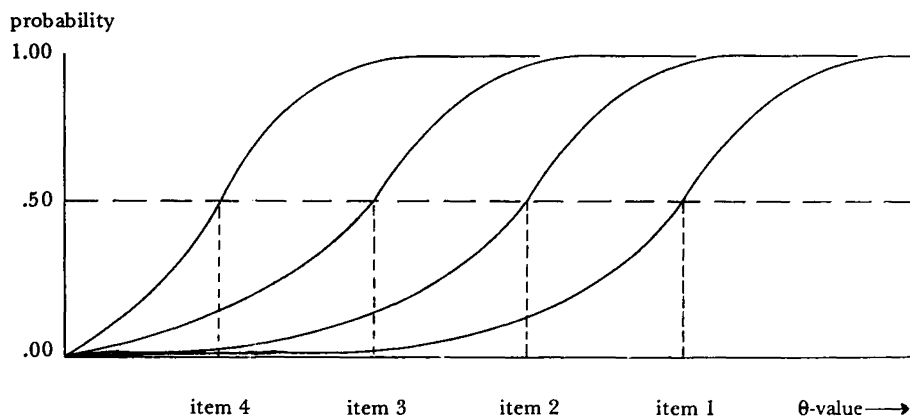
$$m \geq 10; m \geq 5; m \geq 2; m \geq 1$$

All edges of a partial graph at a higher level of multiplicity are contained in a graph at a lower level of multiplicity; the graph at a lower level of multiplicity is obtained by adding a number of edges to the partial graph at a higher level of multiplicity. The graph at a higher level of multiplicity is *nested* in the graph at a lower level.

## 2. Cumulative Scales of Vertices in a Simple Graph

In Figure 3 the stochastic cumulative scaling model of Mokken is given. We are dealing with an underlying continuum  $\theta$  and two sets of elements (in the interview situation: subjects and items). Each subject has an unknown numerical value  $\theta$  on the continuum. For each item we can draw a *trace line*, which gives the probability of a positive response for the different subject values  $\theta$ . Each item also has an unknown numerical value,  $\delta_i$ , on that continuum  $\theta$ . For theoretical reasons we make the value  $\delta_i$  for item  $i$  equal to the value  $\theta$  of that subject that gives the positive response with probability of .5. In the stochastic scaling model of Mokken it is required that the trace lines are doubly monotone or holomorphic: each trace line is monotone non-decreasing for increasing subject values; the different trace line may not intersect (see Figure 3).

Figure 3. Trace Lines for Four Holomorphic Items



We can represent the data on which we apply the Mokken-model, as a bipartite graph. Each vertex in the first subset represents a subject, each vertex in the second subset represents an item. A subset and an item are directly connected by an edge if that subject responds positively to that item. The structure of the data is equal to that of Figure 2, the bipartite graph of sponsorship in the United Nations. It depends on the theoretical perspective whether we consider the delegations as the subset of subjects or as the subset of items. In Stokman and Stokman and Van Schuur the delegations were considered as items and the resolutions as subjects, because we were

interested in cumulative leadership scales of delegations<sup>8</sup>. We will now consider which implications the cumulative scaling model has for the structure of the network between items as induced by the subjects. In case of the sponsorship data we therefore consider which implications a cumulative structure of the relations in the bipartite graph between delegations and resolutions has for the network of co-sponsorship relations between delegations, which was generated from the bipartite graph by induction.

For a monotonely homogeneous set of  $k$  items it can be proven that all items,  $i, j$  ( $i, j = 1, 2, \dots, k$ ) are positively correlated. In Table 2 two items are cross-tabulated, both in terms of frequencies and in terms of probabilities:  $n_i$  gives the frequency ( $\pi_i$  the fraction) of the subjects that respond positively to item  $i$ ;  $n_j$  gives the frequency ( $\pi_j$  the fraction) of the subjects that respond positively to item  $j$ .  $n_{++}$  is the number of subjects ( $\pi_{++}$  the fraction) that respond positively to both item  $i$  and item  $j$ . For  $k$  items to form a cumulative scale it can now be proven that for all pairs of these  $k$  items:

$$\pi_{++} > \pi_i \cdot \pi_j \quad \text{or}$$

$$\frac{n_{++}}{N} > \frac{n_i \cdot n_j}{N \cdot N}$$

Table 2. Cross-Tabulation of Two Items  $i$  and  $j$

(a) in Terms of Frequency				(b) in Terms of Probabilities			
item $j$				item $j$			
item $i$	+	-		+	-		
	$n_{++}$	$n_{+-}$	$n_i$	$\pi_{++}$	$\pi_{+-}$	$\pi_i$	$\pi_i$
	$n_{-+}$	$n_{--}$	$N - n_i$	$\pi_{-+}$	$\pi_{--}$	$1 - \pi_i$	$1 - \pi_i$
	$n_j$	$N - n_j$	$N$	$\pi_j$	$1 - \pi_j$		$1$

The frequency  $n_{++}$  between items  $i$  and  $j$  is equal to the multiplicity of the edges between item  $i$  and  $j$  in the induced network. *Items form a cumulative scale only if the multiplicity of the edges between these items in the network induced by the subjects is larger than that expected in the case of random response*<sup>9</sup>. For the spon-

<sup>8</sup> See Stokman, Roll Calls and Stokman, Frans N., and Van Schuur, Wijbrand H., Leadership Dimensions among Developing Nations in the United Nations (reproduced in this volume).

<sup>9</sup> Mokken uses the coefficient  $\Phi/\Phi_{\max}$  to measure the positive correlation between each pair of items. Loevinger's coefficient of scalability  $H$  is used to judge the scalability of a whole scale. This coefficient can be written as a function on the whole matrix of the  $\Phi/\Phi_{\max}$  coefficients ( $H_{ij}$ 's) between each pair of items (Mokken, 1970, 150). The procedure of multiple scal-

sorship data it implies, that delegations form a cumulative (leadership) scale only if the multiplicity of the edges between these delegations in the network of co-sponsorship relations is larger than that expected in the case of random sponsorship.

Double monotony or *holomorphism* of the items can be checked by inspection of two matrices, one matrix containing the fractions  $\pi_{++}$  of all pairs and another matrix containing the fractions  $\pi_{--}$  of all pairs (see Table 2 (b)). This check for holomorphism is described by Mokken. Let us consider in particular the matrix of the fractions  $\pi_{++}$ , because this matrix corresponds to the network between the items, as induced by the subjects. If the items are ordered according to their difficulty on the basis of the fraction of positive responses ( $i < j : \pi_i \leq \pi_j$ ), the fractions  $\pi_{++}$  in each row of the matrix should increase monotonely with column index  $j$ . This is a necessary condition for a holomorphic set of items. This implies that the highest  $\pi_{++}$  fractions, corresponding with the highest multiplicities  $n_{++}$  in the network between items induced by the subjects, should be observed between the easiest items, *i. e.* the items with the largest number of positive responses. For the sponsorship data it implies, that the highest multiplicities  $n_{++}$  should be observed between the most active delegations. In Section 1 we considered a repeated analysis of a network. In this repeated analysis the minimal levels of multiplicities were decreased with each step. This stepwise procedure works out in a very specific way, if we are dealing with a network between items that form a doubly monotone cumulative scale. In each step the repeated analysis of the network then results in one dense component (and a number of isolated vertices). Moreover, over the consecutive steps with decreasing multiplicities items will be added to the component in the order of decreasing item difficulty: first, the two easiest item form one component; then, the next easiest items will be added to the component and so on until in the last step the most difficult item will be added to the component. *We may therefore conclude that the existence of one component with a high component density ( $a_c$ ) at the different levels of multiplicity in an induced network indicates that the relations are cumulative in the graph that generated that network.*

ing selects items from a set of items in such a way that the scale coefficient  $H$  is maximized. It can be conceived as a cluster procedure on the network between items, as induced by the subjects. However, in this procedure we do not weight the edges according to their multiplicity ( $n_{++}$ ): here we make the weight of an edge between items  $i$  and  $j$  equal to the correlation coefficient  $H_{ij}$ .

### 3. Cumulative Scales of Relations in a Multigraph

The stochastic scaling model of Mokken can also be used to test the cumulative character of relations in a multigraph<sup>10</sup>. Let us consider a network of  $p$  vertices and  $k$  different relations between these  $p$  vertices. We therefore consider  $k$  different graphs, defined on the same set of vertices. The  $k$  graphs can be represented in one multigraph by associating the relation as information with the edge. In the example of the co-sponsorship data we consider the co-sponsorship relations separately for  $k$  issues. We therefore consider  $k$  different graphs of co-sponsorship relations between the  $p$  delegations; two delegations are directly connected by an edge in the  $i^{\text{th}}$  graph, if they co-sponsored one or more proposals on issue  $i$ . The different *relations* can now be considered as the *items* in the scale model; each *pair of vertices* is considered as a respondent or *subject* in the scale model<sup>11</sup>; the *existence of an edge* in the  $i^{\text{th}}$  graph between a pair of vertices is equivalent to a *positive response* of a subject to that item in the scale model. The density in the  $i^{\text{th}}$  graph is related to the difficulty of the  $i^{\text{th}}$  item (the item difficulty) in the scale model. For a holomorphic set of items the order of the fractions  $\pi_i$  in the population (see Table 2) are determined by the order of the item difficulties  $\delta_i$ . The sample fraction  $P_i$  is an unbiased and relatively precise estimate of the population fractions  $\pi_i$ <sup>12</sup>. The density of the  $i^{\text{th}}$  graph is the fraction of pairs of vertices between which the  $i^{\text{th}}$  relation has been observed. *The density of the  $i^{\text{th}}$  graph is therefore equal to the population fraction  $\pi_i$  or can be used as an unbiased and relatively precise estimate of it. The graph theoretical elaboration of the stochastic cumulative scaling model gives the density of the graph therefore a number of desirable conceptual and statistical properties.*

We conclude that the different cumulative dimensions of relations in a multigraph can be determined with the theory and procedures of scale analysis as developed by Mokken.

<sup>10</sup> The applicability of the theory and procedure of scale analysis in multigraphs was elaborated by Mokken and the author in the context of a research project on economic and political power.

<sup>11</sup> In table 2  $N$  is therefore equal to  $1/2 p(p-1)$ , the total number of pairs of vertices in the graph.

<sup>12</sup> Mokken, Theory and Procedure, pp 126–8.

#### 4. Generation and Analysis of the Networks for Each Cumulative Dimension of Relations

Scale analysis of the different relations in a multigraph, discussed in the last section, results in one or more cumulative scales of relations. These scales can now be used to generate networks between the vertices for each cumulative dimension separately. These networks can then be analyzed to determine the structure and central vertices for the different dimensions.

Suppose  $k$  relations formed a cumulative scale over pairs of vertices. In the example of the co-sponsorship data, suppose that co-sponsorship is cumulative over  $k$  issues. We can now determine the *summation score* for each pair of vertices over these  $k$  relations: the number of relations that has been observed between that pair of vertices. In the model of scale analysis this score is known as the score of a response pattern. It is a good estimate of the order of the pairs of vertices (the subjects) on the underlying dimension of relations<sup>13</sup>. *For each cumulative dimension of relations we therefore generate a network in which the weight of the edges is equal to the summation score over the  $k$  relations in the scale.* These networks can then be analyzed with graph theoretical concepts and measures to determine the structure and central vertices.

An example of such an elementary graph analysis of networks for different cumulative dimensions of relations can be found in Stokman<sup>14</sup>. In the period 1960–63 two cumulative dimensions of co-sponsorship relations existed among the developing nations: one of colonial issues and one of socio-economic issues. We therefore generated two networks among the developing nations: a colonial network and a socio-economic network, the weights of the edges being the summation score over the issues in the issue dimension. Particularly the relations within Latin America and those between Latin America and Afro-Asia were quite different for the two networks.

<sup>13</sup> Op. cit., pp. 128–129.

<sup>14</sup> Stokman, Roll Calls.

## Cliques, Clubs and Clans\*

### 1. Introduction

In the analysis of social networks adequate concepts are necessary to indicate various types and configurations of social groups such as peer groups, *coteries*, acquaintance groups, etc. The problem has theoretically been argued convincingly a. o. by Kadushin, who introduced the general concept of „social circle“<sup>1</sup>. In the actual empirical study of social networks there is therefore a need for adequate operational and analytically useful concepts to represent such more or less closely knit groups.

Many of these can be developed with the help of the theory of graphs and networks. A well-known concept, more or less corresponding to that of the peer group is the clique: a group all members of which are in contact with each other or are friends, know each other, etc. However, similar concepts will be necessary to denote less closely knit, yet significantly homogeneous social groups, such as „acquaintance groups“, where every pair of members, if they are not in mutual contact, have mutual acquaintances, or common third contacts, etc. In this latter type of social group an important aspect is given by the question whether the homogeneity of a social group is due to its position in a larger social network in which it is embedded, or whether it is a property of the group itself as a more or less *autarchic* unit, *independent* of the surrounding social network. In the *first* case, for instance, a group may be as closely knit as an „acquaintance network“, just because there are „mutual acquaintances“ outside the group in the surrounding network. Changes in the environment, *i. e.* the outside network, may change the character of the group as an acquaintance group. In the *latter* case, however, a group is an acquaintance group, because mutual acquaintances linking members are themselves members of that group. Therefore changes in the outside social network will not effect the nature and structure of the group itself as an acquaintance group. Such concepts can be

\* Part of this research was performed during the authors stay as Visiting Professor at the University of Michigan. The author owes much to the stimulating discussions with Professor Frank Harary.

<sup>1</sup> Kadushin, Charles, Power, Influences and Social Circles: A new methodology for studying opinion makers, in: *American Sociological Review*, 33 (1968), pp. 685–699.



worked out in terms of graph theoretic cluster concepts. For instance, the familiar concept of *n-clique* deals with the tightness of a group as a *global* property, due to the interrelationships or interactions of *all* members of a larger social network. The concepts of *clubs* and *clans*, to be introduced here denote a *local* property of structural *autarchy* in the sense that the interrelationships within the particular social group are sufficient for its homogeneity, and independent of those interrelationships involving members or parts of the surrounding larger social network.

In this paper we shall introduce three different cluster concepts of graphs — cliques, clubs and clans — and investigate their interrelationships. The graphs treated here will be simple graphs: finite, nonempty, and having no loops or multiple lines. We shall mainly follow the notation and concepts given by Harary, to which we may refer the reader for further reference<sup>2</sup>. We shall be satisfied here with a cursory introduction of the concepts and notation used here.

A graph  $G$  is a set of points together with a set of lines. To simplify notation here, we shall use the same symbol  $G$  to denote the set of points of  $G$ . Any line of  $G$  connects some pair of points  $u, v \in G$ , which then are said to be adjacent to each other in  $G$ . We shall also consider subgraphs of  $G$ , indicated by their pointset. If  $H \subset G$  is a subset of  $G$ , the subgraph  $H$  of  $G$  consists of all points of  $H$  together with all lines of  $G$ , which connect points  $u, v \in H$  in  $G$ . A path, connecting two points  $u, v$  of a subgraph  $H$ , in that same subgraph  $H$ , consists of points  $u, w_1, w_2, \dots, w_{l-1}, v \in H$ , such that  $u$  is adjacent to  $w_1$ ,  $w_1$  is adjacent to  $w_{i+1}$ , consecutively and  $w_{l-1}$  is adjacent to  $v$ . The length  $l$  of a path is given by the number of its lines. A cycle  $C_l$  of length  $l$  is a path of length  $l$ , where  $u = v$ . A subgraph  $H$  is connected in  $G$ , if each pair of points  $u, v \in H$  is connected by a path in  $H$ . A complete graph  $K_p$  is a graph of  $p$  points, where each pair of points is adjacent to each other.

We shall also consider maximal subgraphs with respect to a given property. They are subgraphs of  $G$  satisfying that property, such that no larger subgraphs with that property exist in  $G$ , which contain them. A well known example is given by the cliques of a graph  $G$ : maximal complete subgraphs  $K_p$  of  $G$ .

The distance of a pair of points  $u, v$  in a certain subgraph  $H$ , denoted by

$$d_H(u, v)$$

is given by the length of a shortest path connecting  $u$  and  $v$  in  $H$ .

If  $u, v$  are not connected in  $H$ , this distance is infinity. We shall frequently make use of the well-known relation that, if  $H$  is a subgraph of  $G$ , then for every pair of points  $u, v \in H$

$$d_G(u, v) \leq d_H(u, v) \quad (1)$$

The distance between any two points in a subgraph of  $G$  cannot be smaller than their distance in  $G$  itself.

The diameter of a subgraph  $H$  is given by the largest distance between a pair of points in that subgraph.

<sup>2</sup> Harary, F., Graph Theory, Reading/Mass. 1969.

If we extend the pointset  $H \subset G$  with a point  $w \in G-H$  or a subset  $S \subset G-H$ , we may consider the distances in the larger subgraphs corresponding to  $H \cup \{w\}$  or  $H \cup S$ , denoting for simplicity distances as  $d_{H,w}$  or  $d_{H,S}$ .<sup>\*</sup> The degree of a point  $u$  in  $G$  is the number of points (neighbors) adjacent to  $u$  in  $G$ . The set of those points is called the 1-neighborhood of  $u$  in  $G$ , to be denoted by  $V_1(u)$ . We may restrict the set of neighbors to those in a subgraph  $H$  of  $G$  only, to be denoted as  $V_1^H(u)$ . Similar extensions may be made to  $n$ -neighborhoods of  $u$ : points at distance  $n$  of  $u$ .

## 2. Cluster Concepts

In standard graph theory a familiar cluster concept is given by the *cliques* of a graph  $G$ . As mentioned above, they are given by the set of maximal complete subgraphs of  $G$ . Another cluster type definition of subgraphs of graphs are the „ $n$ -cliques“, introduced by Luce<sup>3</sup> as given by the following definitions.\*\*

*Definition 1.* An  $n$ -clique  $L$  of a graph  $G$  is a maximal subgraph of  $G$  such that for all pairs of points  $u, v$  of  $L$  the distance in  $G$

$$d_G(u, v) \leq n. \quad (2)$$

The reader may note that, due to the maximality of  $L$ , for every point  $w \in G-L$ , there is a point  $v \in L$  for which

$$d_G(w, v) > n \quad (3)$$

From this definition, it can be seen that the  $n$ -clique is a *global* concept, based on the total structure of the network, as based on the graph and reflected in its distance matrix. The distances between points in a certain subset of points, can be based on shortest paths, involving other points from the network not belonging to that group.

It is well-known, therefore, that in the *subgraph*, formed by the points of an  $n$ -clique  $L$ , the distances between points can be larger than  $n$ . This follows from the

<sup>\*</sup>The conventional notation of set theory is used. In particular  $S \subseteq G$  will denote set inclusion,  $S \subset C$  proper inclusion and  $S = G$  identity of sets.

<sup>3</sup> Luce, R.D., Connectivity and Generalized Cliques in Sociometric Group Structure, in: Psychometrika, 15 (1950), pp. 169–190. See also Luce, R. D., and Perry, A., A Method of Matrix Analysis of Group Structure, in: Psychometrika, 14 (1949), pp. 94–116.

\*\* Although the cluster concepts, to be introduced in this paper, suggest more appropriate names, we shall resist the temptation to do so and accept this part of the nomenclature as established.

familiar property, referred to above in (1), that for any two points  $u, v$  of the subgraph  $L$  of  $G$  we must have

$$d_G(u, v) \leq d_L(u, v).$$

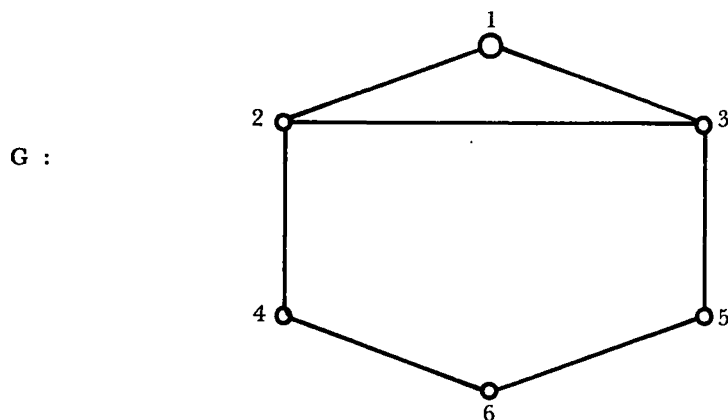
The condition (2) therefore does not imply that for each  $u, v \in L$

$$d_L(u, v) \leq n.$$

Consequently, the diameter of  $L$  may be larger than  $n$ .

In a recent article Alba has illustrated this phenomenon with the example, given here in Figure 1<sup>4</sup>.

Figure 1: Graph  $G$



If we restrict our attention to 2-cliques  $L$  ( $n = 2$ ) and designate 2-cliques by their pointsets, it can be seen that  $L = \{1, 2, 3, 4, 5\}$  is a 2-clique. However, its diameter is 3, *i. e.* the largest distance in  $L$  is 3, in the case of the pair of points 4 and 5. In fact, an  $n$ -clique can be disconnected (diameter infinity), as we shall illustrate further in this paper.

The concept of  $n$ -clique therefore does not embody the idea of particular tightness or even connectedness of the particular group concerned as an essential feature of the corresponding cluster of points in a graph. Yet in many, if not most, problems in social network analysis, leading to a graphtheoretic formulation, this idea of interconnectedness is a basic feature of the „tightness“ of sets of points, underlying the definition of a cluster. Putting up a similar argument for the connectedness of

<sup>4</sup> Alba, R. D., A Graph-Theoretic Definition of a Sociometric Clique, in: Journal of Mathematical Sociology, 3 (1973), pp. 113–126.

clusters as subgraphs, as well as for their tightness as measured by their diameter, Alba introduced „sociometric cliques“, as a more satisfactory subclass of  $n$ -cliques. They are  $n$ -cliques with diameter  $n$  and consequently connected. As the sociometric context is not essential, we suggest as a more appropriate name „ $n$ -clan“.

*Definition 2.* An  $n$ -clan  $M$  of a graph  $G$  is an  $n$ -clique of  $G$  such that for all pairs of points  $u, v$  of  $M$  the distance  $M$

$$d_M(u, v) \leq n. \quad (4)$$

Consequently, for an  $n$ -clan  $M$  of  $G$  the following relations hold:

(1) for all points  $u, v \in M$ :

$$d_M(u, v) \leq n; \quad (5)$$

(2) for all points  $w \in G - M$  there is a  $u \in M$  for which:

$$d_G(u, w) > n. \quad (6)$$

The relations (5) and (6) imply that  $M$  is an  $n$ -clique, as from (1) and (5) we have

$$d_G(u, v) \leq d_M(u, v) \leq n, \quad (7)$$

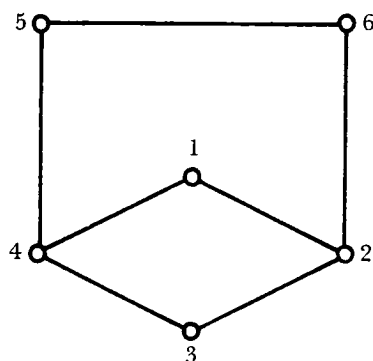
whereas (6) ensures the maximality of  $M$  as an  $n$ -clique.

Instead of the restriction of cliques to clans we might as well have looked immediately at clusters corresponding to subgraphs of diameter  $n$ . These we shall call „ $n$ -clubs“.

*Definition 3.* An  $n$ -club  $N$  of a graph  $G$  is a maximal subgraph of  $G$  of diameter  $n$ . For all points  $u, v$  of an  $n$ -club  $N$  we have for the subgraph  $N$  of  $G$

$$d_N(u, v) \leq n. \quad (8)$$

Figure 2:  $N$  of Diameter 2



$N$  :

The maximality of  $N$  as an  $n$ -club of  $G$  implies that for all points  $w \in G - N$ , there is a point  $u \in N$  such that

$$d_{N,w}(u, w) > n.$$

This condition, however, is not sufficient for the maximality of  $N$ , as illustrated by the graph of Figure 2. The points on the cycle  $C_4: \{1, 2, 3, 4\}$  form a subgraph of diameter 2. Neither point 5 nor point 6 can be added without enlarging the diameter of the resulting subgraph to 3. Yet  $N$  as a whole is a graph of diameter 2 which contains the  $C_4$ .

### 3. Interrelationship of Cliques, Clubs and Clans

By definition  $n$ -clans are  $n$ -cliques of diameter  $n$ . But how are they related to  $n$ -clubs? According to Alba all  $n$ -clubs are  $n$ -clans, as formulated in his theorem 2.1, which, in our terminology states that a subgraph of  $G$  is an  $n$ -clan if and only if (iff) it is an  $n$ -club. This theorem is incorrect, as the „if“-part is deficient.

This can be shown with the following proposition.

*Proposition 1. Every  $n$ -club  $N$  of a graph  $G$  is contained in some  $n$ -clique  $L$  of  $G$ .*

*Proof:* An  $n$ -club  $N$  of  $G$  satisfies (8) and, as a subgraph also (1).

Therefore, we have for all points  $u, v \in N$

$$d_G(u, v) \leq d_N(u, v) \leq n$$

Hence  $N$  is contained in some  $n$ -clique  $L$  of  $G$ . However,  $N$  can be properly contained in such an  $n$ -clique  $L$ . For instance, there may be a point  $w$  of  $L$ , not in  $N$ , such that for all points  $u \in N$  we have

$$d_G(u, w) \leq n,$$

whereas there is a point  $v \in N$  such that

$$d_{N,w}(v, w) > n$$

For instance, in the example provided by Alba, as given in Figure 1 here, the set  $\{1, 2, 3, 4\}$  is obviously a 2-club, which is not a 2-clique and therefore not a 2-clan, as it is properly contained in the 2-clique  $\{1, 2, 3, 4, 5\}$ .

*Proposition 2. Every  $n$ -clan  $M$  of  $G$  is an  $n$ -club of  $G$ .*

*Proof:* Let  $M$  be an  $n$ -clan of  $G$ . It therefore satisfies (5), (6), and (7). Now assume  $M$  to be contained in a larger subgraph of diameter  $n$  formed by  $M \cup S$ , where  $S \subset G - M$ :  $S$  is a subset of points in the other points of  $G$ . However, then we must have, for all  $u, v \in M$ ;  $s, w \in S$ :

$$d_G(u, w) \leq d_{M, S}(u, w) \leq n;$$

$$d_G(s, w) \leq d_{M, S}(s, w) \leq n;$$

and obviously,

$$d_G(u, v) \leq d_{M, S}(u, v) \leq d_M(u, v) \leq n.$$

This violates (6) and contradicts the maximality of  $M$  as an  $n$ -clique of  $G$ :  $M \cup S$  is contained in some  $n$ -clique  $L$  of  $G$ .

Consequently, there can be no such set  $S$  in  $G$  and  $M$  is a maximal subset of  $G$  with diameter  $n$ . That is,  $M$  is also an  $n$ -club of  $G$ .

In our example of Figure 1, the only 2-clique, which is also a 2-clan and hence a 2-club is formed by the set of points  $\{2, 3, 4, 5, 6\}$ .

An obvious corollary of proposition 2 is:

*Corollary 1. If an  $n$ -club  $N$  of  $G$  is contained in an  $n$ -clan  $M$ , then  $N = M$ .*

Our results can be summarily illustrated with the aid of Fig. 1, where we restrict ourselves to distance 2 or diameter 2, ( $n = 2$ ).

a) 2-cliques of  $G$ :

a1:  $\{1, 2, 3, 4, 5\}$ ;

a2:  $\{2, 3, 4, 5, 6\}$ .

b) 2-clubs of  $G$ :

b1:  $\{1, 2, 3, 4\}$ ;

b2:  $\{1, 2, 3, 5\}$ ;

b3:  $\{2, 3, 4, 5, 6\}$ .

c) 2-clans of  $G$ :

c1:  $\{2, 3, 4, 5, 6\}$ .

The 2-clubs b1 and b2 are not 2-clans. They are properly contained in the 2-clique a1, which is *not* a 2-clan, as it has diameter 3. The 2-clique a2 is a 2-clan (c1) and hence also a 2-club (b3).

#### 4. The Systems of Cliques, Clubs and Clans of a Graph

From the foregoing discussion it will be clear that for any graph  $G$  we can distinguish:

- the system of  $n$ -cliques of  $G$ : the class  $\mathcal{L}_n(G) = \mathcal{L}_n$ , the elements of which are indicated by the pointsets  $L$  of the different  $n$ -cliques  $L$  of  $G$ ;
- the system of  $n$ -clubs of  $G$ : the class  $\mathcal{N}_n(G) = \mathcal{N}_n$ , containing the pointsets  $N$  of the different  $n$ -clubs  $N$  of  $G$ ;
- the system of  $n$ -clans of  $G$ : the class  $\mathcal{M}_n(G) = \mathcal{M}_n$ , containing the pointsets  $M$  of the different  $n$ -clans of  $G$ .

In this paragraph we shall consider more closely the possible interrelationships of these classes  $\mathcal{L}_n$ ,  $\mathcal{N}_n$  and  $\mathcal{M}_n$  of a graph  $G$ .

Consider the symmetric difference of  $\mathcal{L}_n$  and  $\mathcal{N}_n$ :

$$\mathcal{L}_n \Delta \mathcal{N}_n = (\mathcal{L}_n \cup \mathcal{N}_n) - (\mathcal{L}_n \cap \mathcal{N}_n),$$

containing only those  $n$ -cliques or  $n$ -clubs which are not common to both. Define

$$\overline{\mathcal{N}}_n \stackrel{d}{=} \mathcal{N}_n \cap (\mathcal{L}_n \Delta \mathcal{N}_n),$$

which contains only  $n$ -clubs which are not  $n$ -clans, and

$$\overline{\mathcal{L}}_n \stackrel{d}{=} \mathcal{L}_n \cap (\mathcal{L}_n \Delta \mathcal{N}_n),$$

the subclass of  $n$ -cliques, which are not  $n$ -clans.

The foregoing results of propositions 1 and 2 can be collected in the following lemma.

*Lemma 1. For each  $n$ -club  $N \in \mathcal{N}_n$ , there is an  $n$ -clique  $L \in \mathcal{L}_n$  such that  $N \subseteq L$ ;*  
*(a)  $N$  is an  $n$ -clan ( $N \in \mathcal{M}_n$ ) iff for every  $v \in G - N$  there is a  $u \in N$  such that*

$$d_G(u, v) > n;$$

*(b)  $N$  is not an  $n$ -clan ( $N \in \overline{\mathcal{N}}_n$ ) iff there is a  $v \in G - N$ , such that for all  $u \in N$*

$$d_G(u, v) \leq n.$$

According to lemma 1 every  $n$ -club  $N \in \mathcal{N}_n$  is either equal to an  $n$ -clique  $L \in \mathcal{L}_n$ , ( $N=L$ ), and then an  $n$ -clan ( $N \in \mathcal{M}_n$ ) or  $N$  is properly contained in some  $n$ -clique  $L \in \mathcal{L}_n$ , ( $N \subset L$ ). The concept of  $n$ -clique ( $L \in \mathcal{L}_n$ ) defines a class of clusters or subgraphs based on „close“ reachability of points, through paths including points in  $G$  external to  $L$ .

On the other hand  $n$ -clubs ( $N \in \mathcal{N}_n$ ) are based on the condition of „close“ reachability of points, involving *internal* points of  $N$  only. In that sense  $n$ -clubs  $N$

have a property of *local autarchy*: the closeness or tightness of their communication structure is independent of the relations of its points with the surrounding network. Obviously, this latter, more stringent, condition leads to „smaller“ clusters: n-clubs cannot be larger than n-cliques, as they are included in them. In fact, as we can see from the examples mentioned in this paper, an n-clique L can contain more than one n-club N and, conversely, an n-club can be contained in more than one n-clique L.

The n-clans ( $M \in \mathcal{M}_n$ ) belong to both  $\mathcal{L}_n$  and  $\mathcal{N}_n$ ; they are n-cliques as well as n-clubs. As n-clubs they share the property of connectedness with sufficiently narrow diameter. As n-cliques they have the advantage of „size“: they are as „large“ as n-cliques.

We therefore can subsume the interrelationship of these classes in the following three, mutually exclusive subclasses:

- a)  $\mathcal{M}_n = \mathcal{L}_n \cap \mathcal{N}_n$ : the class of n-clans as the intersection of the class of n-cliques and the class of n-clubs;
- b)  $\overline{\mathcal{L}}_n$ : the subclass of n-cliques which are not n-clubs;
- c)  $\overline{\mathcal{N}}_n$ : the subclass of n-clubs, which are properly contained in n-cliques.

For  $n = 1$  we trivially have  $\mathcal{L}_n = \mathcal{N}_n = \mathcal{M}_n$ , all systems reducing to the system of cliques of G: the class of maximal complete subgraphs of G. A similar trivial reduction can be seen for the case of G itself being an n-club. Excluding these trivial cases, we may note that, except for the nullgraph,  $\mathcal{N}_n$  and  $\mathcal{L}_n$  are never empty.

The following three cases deserve some interest:

- I.  $\mathcal{M}_n = \phi$ : there are no n-clans;
- II.  $\overline{\mathcal{N}}_n = \phi$ : all n-clubs are n-clans; (Alba's case);
- III.  $\overline{\mathcal{L}}_n = \phi$ : all n-cliques are n-clans.

*Case I.*  $\mathcal{M}_n = \phi$ . (No clans). In this case  $\overline{\mathcal{N}}_n$  and  $\mathcal{L}_n$  are disjoint classes. Every  $N \in \overline{\mathcal{N}}_n$ , so condition (b) of lemma 1 holds for every  $N \in \mathcal{N}_n$ . We shall illustrate this case for  $n = 2$  (distances and diameter 2). Extensions to general n, if necessary, are left to the reader.

We can define for all  $N \in \overline{\mathcal{N}}_2 (= \mathcal{N}_2$  in this case) and for every  $v \in G - N (\neq \phi)$  the set

$$N_v \stackrel{d}{=} \{u \in N; d_{N,v}(u, v) > 2\} \quad (9)$$

Note that  $N_v$  is never empty, as  $G - N$  never is, under the present assumptions. Consequently, for every  $v \in G - N$  and for all corresponding  $u \in N_v$  we have

$$V_1^{(N)}(u) \cap V_1^{(N)}(v) = \phi \quad (10)$$

Their 1-neighborhoods in N are disjoint in G. Obviously,  $v \in G - N$  and  $u \in N_v$  cannot be adjacent in G either, nor in any subgraph of G. Given relation (10) such pair of points u, v can therefore have

$$d_G(u, v) = 2$$



if and only if

$$V_1^{G-N}(u) \cap V_1^{G-N}(v) \neq \emptyset \quad (11)$$

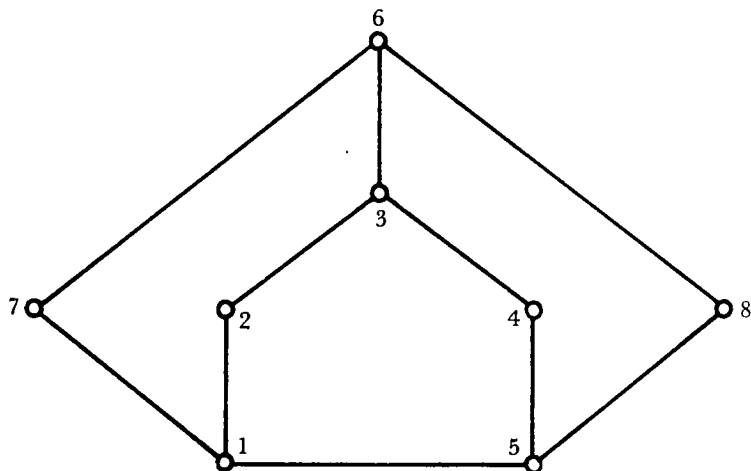
that is when their 1-neighborhoods in  $G-N$  are not disjoint in  $G$ . These considerations establish the validity of the following proposition:

*Proposition 3.*  $\mathcal{M}_2 = \emptyset$  iff for all  $N \in \mathcal{N}_2$ , there is a point  $v \in G-N$ , such that for all  $u \in N_v$

$$V_1^{(G-N)}(u) \cap V_1^{(G-N)}(v) \neq \emptyset$$

Figure 3: A Graph  $G$  Without 2-clans

$G$ :



In this case, as  $\mathcal{M}_2 = \emptyset$ , there are no 2-clans, all 2-clubs  $N \in \mathcal{N}_2$ , are properly contained in larger 2-cliques  $L$ .

This situation is illustrated with the graph  $G$  of Figure 3, which has no 2-clans as can be deduced from the following enumeration of its classes  $\mathcal{N}_2$  and  $\mathcal{L}_n$ :

$\mathcal{N}_2$		$\mathcal{L}_2$
star: {2, 3, 4, 6}	$\subset$	{1, 2, 3, 4, 5, 6}
$C_5$ : {1, 2, 3, 4, 5}	$\subset$	{1, 2, 3, 4, 5, 6}
star: {1, 2, 5, 7}	$\subset$	{1, 2, 3, 5, 6, 7}
$C_5$ : {1, 2, 3, 6, 7}	$\subset$	{1, 2, 3, 5, 6, 7}
star: {1, 4, 5, 8}	$\subset$	{1, 3, 4, 5, 6, 8}
$C_5$ : {3, 4, 5, 6, 8}	$\subset$	{1, 3, 4, 5, 6, 8}
star: {3, 6, 7, 8}	$\subset$	{1, 3, 5, 6, 7, 8}
$C_5$ : {1, 5, 6, 7, 8}	$\subset$	{1, 3, 5, 6, 7, 8}

In this example all 2-cliques contain 2-clubs. More than one 2-club, a star and a cycle  $C_5$ , is contained in each 2-clique.

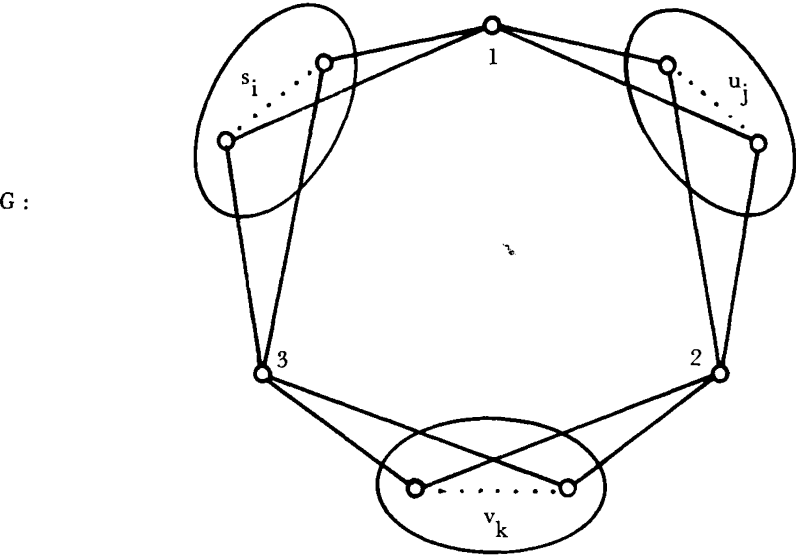
Case II.  $\overline{\mathcal{N}}_n = \phi$ . Alba's case: all clubs are clans. We have  $\mathcal{N}_n \subseteq \mathcal{L}_n$ .<sup>\*</sup> This case is therefore equivalent to  $\mathcal{M}_n = \mathcal{N}_n$ . All  $N \in \mathcal{N}_n$  satisfy condition (a) of lemma 1. For the case of diameter 2 ( $n = 2$ ) one can see easily that:  
 $\mathcal{M}_2 = \mathcal{N}_2$  iff for every  $N \in \mathcal{N}_2$  and for all  $v \in G - N$ , there is a  $u \in N_v$  satisfying  
 (a) 1-neighborhoods in  $G - N$  disjoint:

$$V_1^{(G-N)}(u) \cap V_1^{(G-N)}(v) = \phi \tag{12}$$

or, equivalently,  
 (b) 1-neighborhoods in  $G$  disjoint:

$$V_1(u) \cap V_1(v) \neq \phi \tag{13}$$

Figure 4. A Graph G, All 2-clubs Are 2-clans



An example for  $n = 2$  is given in Fig. 4, where the sets  $\{s_i\}$ ,  $\{u_j\}$  and  $\{v_k\}$  indicate points of degree 2, each adjacent solely to the points 1, 3 or, respectively 1, 2 or 2, 3. All the 2-clubs are 2-clans and therefore 2-cliques. They are:

<sup>\*</sup>It should be noted that, although for each  $N \in \mathcal{N}_n$  we have  $N \subseteq L$  for some  $L \in \mathcal{L}_n$ , in general  $\mathcal{N}_n \subseteq \mathcal{L}_n$  does not hold, as some elements  $N (\in \mathcal{N}_n)$  are not  $n$ -cliques  $L$  and therefore do not belong to  $\mathcal{L}_n$ .

$\{1, 3, s_i\}, \{1, 2, u_j\}, \{2, 3, v_k\}$  and

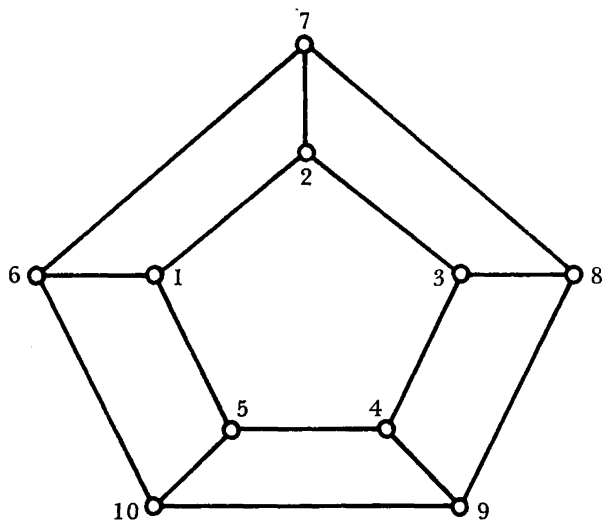
$\{s_i, 1, u_j\}, \{u_j, 2, v_k\}, \{s_i, 3, v_k\}$

There are two 2-cliques, which are not 2-clans:  $L \in \bar{\mathcal{L}}_2$ .

They are:  $\{1, 2, 3\}$  and  $\{s_i, u_j, v_k\}$ . Note that as subgraphs they are nullgraphs *i.e.* totally disconnected. All their points are isolated. It should be noted, that these latter 2-cliques do not contain any 2-clubs, or, for that matter, any subgraph of  $G$  of diameter 2. This illustrates the more general situation where there can exist  $n$ -cliques  $L \in \bar{\mathcal{L}}_n$ , which contain no  $n$ -clubs  $N \in \mathcal{N}_n$ , but at most parts of  $N \in \mathcal{N}_n$ .

*Case III.*  $\bar{\mathcal{L}}_n = \phi$ . All  $n$ -cliques are  $n$ -clans. As  $\mathcal{L}_n \subset \mathcal{N}_n$  we immediately have  $\mathcal{L}_n = \mathcal{N}_n = \mathcal{M}_n$  and therefore  $\bar{\mathcal{L}}_n = \phi$  implies  $\bar{\mathcal{N}}_n = \phi$ . If there are no  $n$ -cliques, which are not  $n$ -clans, then there are no  $n$ -clubs, which are not  $n$ -clans. (We have seen in the former case that the reverse need not be true). It is hard to characterize such graphs beyond the obvious statement, that all  $n$ -cliques have diameter  $n$ .

Figure 5: A Graph  $G$ , All 2-Cliques Are 2-Clans



For  $n = 2$  an example is given by Figure 5.

In that graph all 2-clubs are 2-cliques and conversely. All three clustersystems coincide as  $\mathcal{L}_2 = \mathcal{N}_2 = \mathcal{M}_2$ .

The common elements of these systems are:

- all the stars of degree 3 : *e.g.*  $\{1, 2, 3, 7\}$  etc .
- all the cycles  $C_4$  : *e.g.*  $\{1, 2, 6, 7\}$  etc .
- the two cycles  $C_5$  :  $\{1, 2, 3, 4, 5\}$  and  $\{6, 7, 8, 9, 10\}$ .

## 5. Conclusions and Suggestions for Further Research

We may conclude that the two classes of  $n$ -cliques  $\mathcal{L}_n$  and  $n$ -clubs  $\mathcal{N}_n$  of a graph  $G$  are classes of clusters which are in general but loosely interrelated and have a significance of their own. The latter, the  $n$ -clubs, are maximal subgraphs  $N$  of  $G$  with respect to *internal* reachability of points within distance  $n$ , *i. e.* independent of the connection of the points of  $N$  with other points in  $G : G-N$ . In that sense  $n$ -clubs are essentially *local* concepts: their reachability as diameter  $n$  subgraphs is not effected by changes in the subgraph  $G-N$  and the connection of  $G-N$  with  $N$ . This independence of the environment given by the outer network  $G-N$  can be seen as a certain *local autarchy*. In short,  $N$  as a subgraph would have at most diameter  $n$  in any other graph  $G$ .

The  $n$ -cliques  $L$  are *global* concepts in  $G$  in the sense that their reachability of points within distance  $n$  can involve points *external* to  $L$ . Hence their reachability can be determined outside  $L$  in  $G-L$ : elimination of points from  $G-L$ , or lines in the subgraph  $G-L$  or connecting  $G-L$  and  $L$  can effect the reachability of points in  $L$ .

The  $n$ -clans  $M$  of  $G$ , when they exist, combine these local and global aspects as they are cliques as well as clubs. However, the class  $\mathcal{M}_n$  of  $n$ -clans may be empty for a graph  $G$ .

Finally,  $n$ -clubs  $N$  are always contained in some  $n$ -clique  $L$ . In that sense they are 'smaller' than  $n$ -cliques. Only  $n$ -clans, as  $n$ -clubs have the 'size' of an  $n$ -clique. Moreover different  $n$ -clubs can be contained in the same  $n$ -clique and different  $n$ -cliques can contain the same  $n$ -clubs. The interrelationship of the three classes  $\mathcal{L}_n$ ,  $\mathcal{N}_n$  and  $\mathcal{M}_n$  therefore can be manifold. One perspective for further research is therefore to characterize graphs  $G$  according to the nature of that relationship. In the cases I, II and III, given in the last paragraph our characterization hardly proceeded beyond that provided by the definitions.

Then the development of adequate (computer-)algorithms for the production of the systems of  $n$ -cliques  $L$ ,  $n$ -clans  $M$  and  $n$ -clubs  $N$  of any graph  $G$  invites further research. The problem is satisfactorily solved for the detection of  $n$ -cliques. A well-known method is given by Auguston and Minker<sup>5</sup>. A reputedly faster algorithm than that referred to by them was recently published by Bron and Kerbosch<sup>6</sup>. Therefore, the problem of detecting the system of  $n$ -clans of a graph reduces to sorting out the  $n$ -cliques of diameter  $n$  from the  $n$ -cliques of that graph. (Alba, 1973). The development of an algorithm for the detection of the system of  $n$ -clubs

<sup>5</sup> Auguston, J. G., and Minker, J., An Analysis of Some Graph Theoretical Cluster Techniques, in: Journal of the ACM, 17 (1970), pp. 571-588.

<sup>6</sup> Bron, C., and Kerbosch, J., Finding All Cliques of an Undirected Graph. (H) (Algorithm 457), in: Communications of the ACM, 16 (1973), pp. 575-577.

of a graph may well be more cumbersome. Our first cursory assessment of this problem suggests that it may be of the order of enumerating the subgraphs corresponding to all subsets of points within the  $n$ -cliques of a graph.

Further research may also concern possible generalizations. One obvious generalization is that of  $(m, n)$ -clans ( $m \geq n$ ) of a graph:  $n$ -cliques which are  $m$ -clubs. Another generalization extends these concepts to directed graphs, with the introduction of directed cliques, clubs and clans.

*[First published in:  
Quality and Quantity 13 (1979), 161-173]*

## VII. Structure of Social Bookkeeping: Problems with the Generation of Complex Data Files



## The Nature of Process-Produced Data — Towards a Social-Scientific Source Criticism

### 1. Introduction<sup>1</sup>

Everyday life is increasingly reflected in administrative bookkeeping systems. This situation is the consequence of the enlargement of municipal functions and the creation of urban and regional information systems within the overall development of the welfare state. The scope of administrative functions is becoming more and more diverse<sup>2</sup>. At the same time more people have become clients of a broad range of formal organizations. The penetration of society through a network of formal organizations has increased considerably<sup>3</sup>.

For the bureaucratic handling of these administrative functions, problems that might be regarded as diffuse in everyday life, must be translated into functionally specific terms. As a result of functional specialization within urban bureaucracies this translation almost always implies the reduction of private contingencies into

<sup>1</sup> The research reported here was supported by funds granted to the Institute for Applied Social Research, University of Cologne, Project: Information Systems and Information Behavior, by the Bundesministerium für Forschung und Technologie/Institut für Dokumentationswesen (IDA 0103). Earlier versions or parts thereof were presented at the QUANTUM-SSHA conference in August 1977 as well as at the 9th World Congress of Sociology, 1978.

Further results of the research are described in: Bick, Wolfgang and Müller, Paul J., *Die Buchführung der Verwaltungen als sozialwissenschaftliche Datenbasis*, in: Müller, Paul J. (ed.), *Die Analyse prozeß-produzierter Daten*, HSF, Vol. 2, Stuttgart 1977, pp. 42–88; Bick, Wolfgang and Müller, Paul J., *Stable Patterns within a Network of Urban Bureaucracies — Domains or Positions?*, paper presented at the 1978 American Sociological Association Meeting, San Francisco. The final report on the project is available as Bick, Wolfgang and Paul J. Müller, *Informationssysteme und Informationsverhalten — Soziologische Grundlagenforschung für eine Informationspolitik*, BMFT FB ID 79–01, Fachinformationszentrum Energie, Mathematik, Physik, Kernforschungszentrum, Eggenstein-Leopoldshafen 1979.

<sup>2</sup> See e. g. Liebert, Roland J., *Disintegration and Political Action, The Changing Functions of City Governments in America*, New York 1976.

<sup>3</sup> See Bick, Wolfgang and Müller, Paul J., *Die Buchführung der Verwaltungen als sozialwissenschaftliche Datenbasis*, in: Müller, Paul J. (ed.), *Die Analyse prozeß-produzierter Daten*, HSF, Vol. 2, Stuttgart 1977, pp. 42–88.



causal models underlying the decision programs of formal organizations. This reduction process leads to a highly selective and biased representation of everyday life. Everyday life is not just neutrally translated but transformed into administrative categories.

The proliferation of formal organizations within industrial societies has led to an increase in the volume of bookkeeping on social behavior. Social bookkeeping data<sup>4</sup> are often process-produced. These are data which accrue as „by-products“ or traces of the daily routines of formal organizations and which are therefore not collected for the purposes of scientific or statistical evaluation<sup>5</sup>.

As the significance of such process-produced data for sociological research grows, it becomes increasingly important to study the representative nature of administrative bookkeeping, and to delineate the kinds of approaches that seem most promising with the use of these kinds of data. Both issues will be dealt with in this essay.

When data are originally collected for purposes other than those of scientific research, there is an inherent danger of vulnerability that occurs when distortions and the context in which the data were collected are unknown. For many years the technology of the sample survey has been regarded as the main means of social science data collection. This longevity has made it possible to determine – on the basis of the method's intrinsic quality control procedures – the limits of the validity and reliability of the collected material.

In the case of process-produced data, the development of a source criticism analogous to „historical source criticism“<sup>6</sup> becomes essential, not only for files relating

<sup>4</sup> See e. g.: Dibble, Vernon K., Four Types of Inference from Documents to Events, in: *History and Theory*, 3 (1963), pp. 203–221.

<sup>5</sup> Process-produced data were first – although differently – defined by Stein Rokkan, see: Rokkan, Stein, *Data Services in Europe*, in: *American Behavioral Scientist*, Vol. 19, 4 (1976), pp. 443–454. In the meantime various associations developed definitions of process-produced data similar to our definition. *QUANTUM* (Program 1975) stressed to point that process-produced data are not collected within scientific data collection routines, but are the products of administrative record-keeping. Within IASSIST the question whether data produced by statistical bureaux should be included is still an open one.

For the perspective within *QUANTUM* see: Müller, Paul J., *IASSIST Newsletter*, Vol. 1, No. 2, February 1977, pp. 17–21.

For the less methodological and more pragmatic oriented discussion see: *IASSIST Newsletter*, Vol. 1, No. 2, February 1977, pp. 13–15.

<sup>6</sup> See e. g.: Bernheim, E., *Lehrbuch der historischen Methode und der Geschichtsphilosophie*, Leipzig 1908; Blankenburg, Erhard (ed.), *Empirische Rechtssoziologie*, München 1975; Derlien, Hans Ulrich, *Methodische Probleme der empirischen Verwaltungsforschung*, Bonn 1978; Ivanov, Kristo, *Quality-Control of Information: On the Concept of Accuracy of Information in Data-Banks and in Management Information Systems*, National Technical Information Service, U.S. Department of Commerce 1972; Murphey, Murray G., *Our Knowledge of the Historical Past*, Indianapolis, Ind., 1973; Narroll, Raoull, *Data Quality Control*, New York 1970; Steffen, Wiebke, *Grenzen und Möglichkeiten der Verwendung von Strafakten als Grundlage kriminologischer Forschung*, in: Müller, Paul J. (ed.), *Die Analyse prozeß-produzierter Daten*, HSF, Vol. 2,

to decision-making processes, but for the compilation of comprehensive and uniform files of data on individuals. The first step in the direction of a „social-scientific source criticism“ is an examination of the quality of representation of everyday life provided by administrative bookkeeping systems.

Administrative bookkeeping data are mainly collected in a standardized fashion, that is through application or record forms. In the encounters between clients and administrations using standardized information collection schemes the process of translating everyday life into administrative categories becomes problematic. The clients themselves are involved in this process and the successful translation therefore depends on their willingness and competence to reduce their individual problems into the tight and narrow information needs of urban bureaucracies. Distortions of the representation of everyday life can therefore be assumed.

Administrative bookkeeping data are the results of translating individual problems and private contingencies into the standardized information collection schemes of public administrations. For analysing the quality of process-produced data we used a multi-level and multi-method approach. The population and the network of institutions were the different levels of analysis. The different instruments were *first*, the analysis of application forms; *second*, surveying clients as well as staff members of organizations; and *third*, observing the translation process involved in the encounter between clients and bureaucracies. In the following sections the characteristic features of administrative bookkeeping data and their inherent quality problems are discussed according to the following research questions.

In *section 2*, the application forms for compiling person-related data which are used in the local setting of Cologne are analysed. We compare the information needs of different administrative tasks and show which information is regarded as relevant and which aspects of everyday life are blocked out in administrative bookkeeping. The guiding research questions are then: How selectively, how fragmentarily are the characteristics of the clients entered into the records of formal organizations? To what extent are the social environments (e. g. family, household, friends) of the clients represented?

In *section 3* we analyse whether the selectivity of representation within single offices can be overcome by interlocking bookkeeping systems. Perceiving administrations as a system we identify the holes within administrative bookkeeping that are consequences of blocking out certain aspects of everyday life.

*Section 4* deals with the analysis of the interorganizational network of information exchange in an urban setting. We applied network analysis to show how urban bureaucracies exchange information through direct communication or other forms of interorganizational linkages. To what extent does the selectivity of exchange of information between formal organizations entail a vulnerability of the network of institutions, and in what way does this vulnerability affect their records?

Stuttgart 1977, pp. 89–108; Tilly, Charles, Clio and Minerva, in: John C. McKinney and Edward Tiryakian (eds.), *Theoretical Sociology: Perspectives and Development*, New York 1970, pp. 434–466; von Brandt, A., *Werkzeuge des Historikers*, 7th ed., Stuttgart 1973.

In *section 5* we focus on evaluation of the records made by administrative personnel. How unambiguous, how reliable are the records which are compiled by administrations, for the producers of the records themselves?

*Section 6* describes our observation of the interaction between clients and bureaucracies. We consider the extent to which personal connotations are brought into the process of „skeletonizing“ peoples' problems. Thus we attempt to demonstrate the differences between administrative view points implicit in their problem solving processes, on the one hand, and peoples' ways of thinking about their own situation, on the other. Further on we asked the clients how they evaluated the information collection process. To what extent can people accept the skeletonizing process within administrative data collection? The guiding research questions were: to what extent do the collected data only tell something about the data collectors themselves? Are administrative bookkeeping data artifacts which tell little about the represented everyday lives?

## 2. Representing Everyday Life within Administrative Bookkeeping

### *2. 1. Administrative Theories of Everyday Life*

Policy goals are only broadly defined in laws and thereby transformed into a set of concepts or ways of thinking about reality. For the standardized and routinized execution of these laws a transformation of the concepts into decision programs is necessary. This transformation is heavily characterized by taking over parts of the already existing operationalizations, thus relying on previously defined categories or public statutes. Administrative decision programs are based on implicit causal models about which information is necessary to collect from the client and which not. These implicit causal models define, a priori, contingencies between variables as well as joint thresholds of variables that make people eligible or accessible for administrative action. Information about clients is therefore collected for the purpose of classifying people and not for finding or testing contingencies operating in everyday life. These implicit causal models determine the points of view that organizations have towards their environment. These causal models or points of view are built around substantive focuses that are of interest to the administrative decision programs.

In comparing different models within administrative decision programs one might think of two rather different structures. Decision programs could be grouped according to the administrative domains to which they belong. Causal models aiming at the same broad concept like „improving the health conditions“ would then

be characterized as having an identical set of required information. But we do not expect that the structure of causal models would lead to segmented information bases. Disjoint sets of required information would, in fact, totally exclude the contingencies existing between such concepts as e. g. „health condition“ and „work condition“. In addition to that, such a model would neglect the existent interrelations between the decision programs of different administrations in different domains.

Therefore we perceive the structure of causal models as a system of unique, but interlocked information profiles relevant to each office. The causal model would be focused on a target variable, i. e. the variable measuring the state of affairs the administration is supposed to influence. The causal model then would include all other variables that are regarded as influencing the target variable (see *Figure 1*) leading to an unique profile of relevant information.

The causal models should build up an interlocked structure for two reasons. First, administrations very often regard decisions reached within one decision program (output) as relevant to the operation of another program (input). These interconnections between programs form hierarchical as well as relational structures: eligibility for a service is often dependent on the possession of a certain licence applied for earlier. And, some services are seen as mutually exclusive with the effect that e. g. the amount of money received within one program is taken into account within another program. Second, administrations are confronted with changes in their environment and are dependent on the information given to them by the population. By using information that is produced by other organizations they increase their information input as well as the supposed quality of the information. Therefore, administrative decision programs build up a system of interlocked decision programs.

Implicit causal models are not only characterized by a certain set of variables that is regarded as explanatory for the focus of administrative concern, but are also based on theories about what parts of the applicant's interpersonal network have to be taken into account. It is assumed that the kinds of social environments which can be regarded in administrative bookkeeping would mainly be those that are legally defined. The definitions of relatives or household members meet this requirement: Relatives are defined by the register office, the members of the household are legally defined by the registration office. Although both dimensions, kinship and co-residence, might vary substantively, we do not expect that friendship or acquaintanceship networks — even if they are much more important in everyday life — will be considered as relevant to describing the social relations of the individual applicant. If these relationships were taken into account the administration would be confronted with the fact that there are no other formal organizations which routinely process such categories. Thus, administrations would have to rely solely on the information provided by individual applicants.

Both, the implicit theories about what information about the individual must be collected for processing a case, and those about what social relationship are regarded as influencing the situation of the individual applicant are dimensions of organi-

Figure 1 A  
The causal model of the rental assistance program

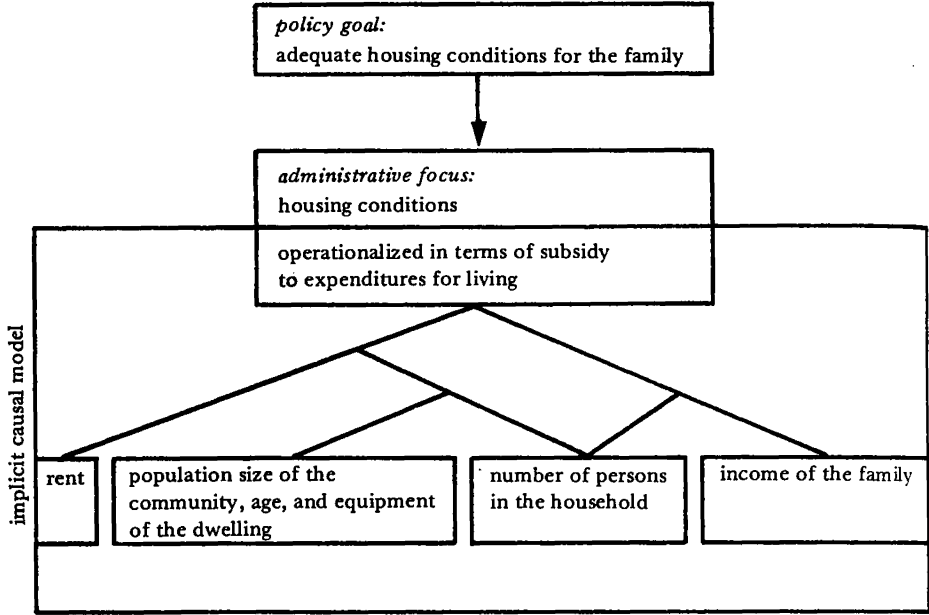
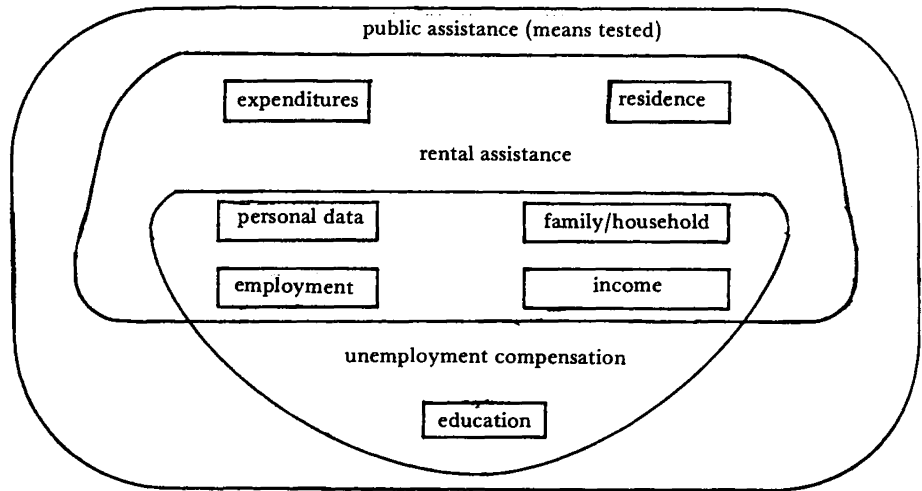


Figure 1 B  
The causal models of three different administrative programs



zational orientations toward clients. They describe the interest of the organizations in the clients' „biographical space“ ranging from a focus upon limited aspects of individual actors to a broad interest in clients and their social environments<sup>7</sup>. These client models determine the biases in the representation of everyday life by formal organizations.

## *2.2. Selectivity and Contingencies within Administrative Causal Models*

Application forms are the operationalizations of causal models within administrative decision programs. They purposively neglect information which is regarded as not relevant by not asking for that information. They seek only those contingencies which are regarded as relevant to the case, and they reflect the interlocked decision programs by asking for multiple statuses of the applicants.

We have analysed a total of 92 application and record forms which are used by administrative offices in the Cologne area for gathering information (see *Appendix 1*). The following administrative domains were included in our study: Cologne city administration, the local labour office in Cologne, and the local health insurance administration. Completed files were not used for our study. Instead, we analysed the standardized interview sheets (applications, record forms, questionnaires, etc.) which are used to compile information for client-oriented administrative tasks. This means that the universe of our analyses is the administrative bookkeeping which compiles person-related data in a standardized fashion.

To test the hypothesis that the causal models of administrations build a system of unique, but interlocked information profiles relevant to the offices we coded the questions in the application forms according to a list of 85 variables (see *Appendix 2*) and analysed the resulting matrix using cluster analytical measures. We first measured the similarity between the information-gathering profiles of administra-

<sup>7</sup> Lefton, Mark and William R. Rosengren, *Organizations and Clients: Lateral and Longitudinal Dimensions*, in: *American Sociological Review*, Vol. 31 (1966), pp. 802–910.

In the text the aggregations of single administrative programs or offices to administrations or domains are described in the figures. Only the various groupings of public assistance and work-related public transfers need special definitions: The program „Sozialhilfe“ is translated into „public assistance (means tested)“, the office providing this assistance is termed „welfare office“.

If the programs of public assistance are combined with work-related public transfers (e. g. benefits from unemployment insurance) the aggregate of the offices providing these types of benefits are termed „social welfare administration“. Finally, if we only refer to the aggregation of means tested assistance programs delivered by the offices of the city administration we use the term „welfare administration“.

tive tasks by means of the non-metric coefficient (Canberra-metric)<sup>8</sup>. This asymmetrical and weighted dissimilarity measure ensures recognition of the similarity of those tasks which are distinguished by a high degree of agreement in terms of the causal models applied. The grouping procedure used, „complete linkage“, was chosen to detect the most homogeneous clusters possible<sup>9</sup>. Figure 2 presents the results of this analysis. It was not possible to distinguish groups of administrative tasks by distinct patterns of information representation. There were no groups found which could be characterized as „natural“ classes in terms of internal homogeneity and external isolation<sup>10</sup>. This means that the representations of the 92 administrative tasks analysed are nearly unique in terms of the respective compilations of certain combinations of attributes. In other words, the administrative causal models cannot be grouped according to administrative domains but instead reflect individual points of view which overlap only partially<sup>11</sup>.

If causal models within administrative decision programs can be characterized as highly selective in regard to the contingencies with which they are concerned, it is still to be expected that certain variables will almost always be collected together with other variables, i. e. that dominant points of view exist. To identify these dominant points of view we aggregated the 85 variables into 15 life sectors and analysed their contingencies. The contingencies between the life sectors „family/household“, „residence“, „employment“, „education/qualifications“, „income“, „expen-

<sup>8</sup> This and the following cluster analyses were done with the program package CLUSTAN 1 C from David Wishart. (Wishart, David, CLUSTAN 1 C User Manual, London 1975.) The „non-metric coefficient“ D (subroutine CORREL, coefficient 37) is defined as

$$D = \frac{b + c}{2a + b + c}$$

(b resp. c are the number of 0/1 resp. 1/0-dismatches, a resp. d the number of 0/0 resp. 1/1-matches).

This coefficient is equivalent to the Canberra-metric for dichotomous variables belonging to the group „Manhattan metric“. See: Lance, T. N., and Williams, N. T., Mixed-Data Classificatory Programs I: Agglomerative Systems, in: Australian Computer Journal, 1 (1967), pp. 15–20.

The non-metric coefficient weights the joint occurrence of attributes (1/1-matches).

<sup>9</sup> The used fusion procedure „complete linkage“ (subroutine HIERARCHY, coefficient 2) defines the distance between two groups as the distance between the furthest elements of the two groups, that is: for a given dissimilarity measure S the distance between X and Y is defined as

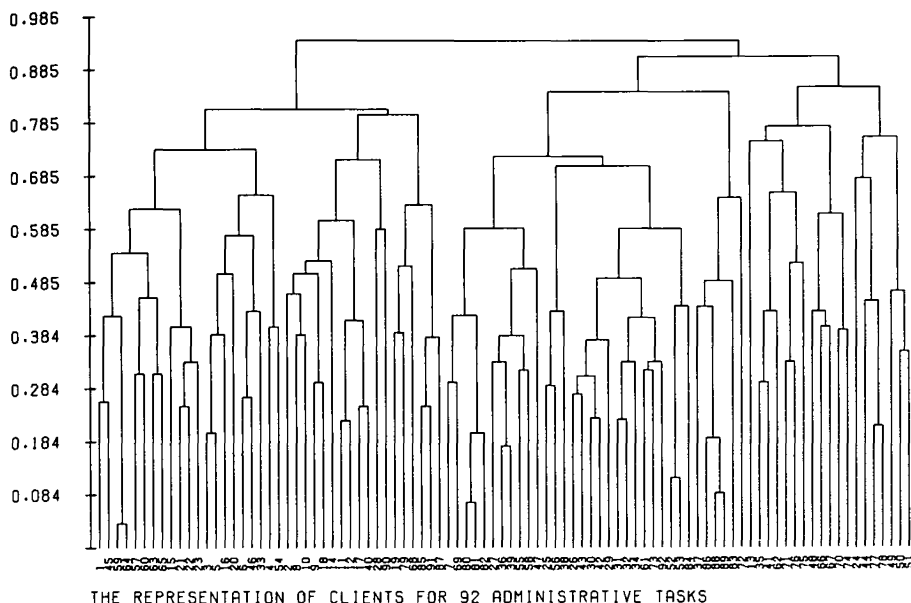
$$S_{X,Y} = \max_{x \in X, y \in Y} S_{x,y}$$

In the hierarchical fusion process those groups/elements are combined which have the highest similarity. As to the advantages of the complete linkage process see: Vogel, Friedrich, Probleme und Verfahren der numerischen Klassifikation, Göttingen 1975, pp. 300 passim.

<sup>10</sup> For the term „natural classes“ see: Sodeur, Wolfgang, Empirische Verfahren zur Klassifikation, Stuttgart 1974, pp. 119 passim.

<sup>11</sup> This result was confirmed by analyses using single and average linkage procedures.

Figure 2



ditures“ and „health“ within the causal models of 92 administrative tasks were measured by means of the coefficient of association<sup>12</sup>. Because it was to be expected that certain life sectors would be represented in varying combinations, approaches which break down these contingencies into disjoint combinations were considered as not being adequate. For this reason we made use of a clustering procedure to ascertain overlapping combinations of interdependent life sectors<sup>13</sup>.

<sup>12</sup>  $\phi$  was computed using the SPSS-routine PEARSON CORR and was then transformed to a similarity measure with a range of 0 to 1.

<sup>13</sup> For this analysis the subroutine KDEND within CLUSTAN 1 C was used. KDEND computes according to the approach developed by Jardine and Sibson, for any given similarity threshold S and number of maximally allowed overlap M, those groups which consist of elements having a similarity equal or higher than S and which have no more than M elements in common. For the approach see: Jardine, N., and Sibson, R., *The Construction Journal*, 11 (1968), pp. 48–50.

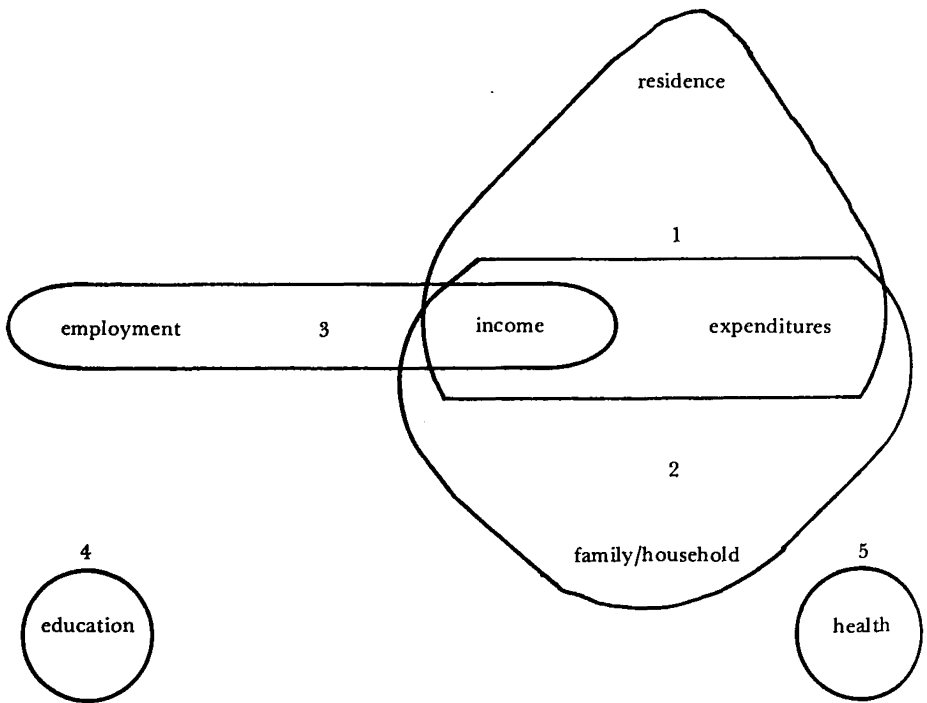
Cole, A. J., and Wishart, D., *An Improved Algorithm for the Jardine-Sibson Method of Generating Overlapping Clusters*, in: *The Computer Journal*, 13 (1970), pp. 156–163.



Figure 3 shows the overlapping contingencies<sup>14</sup>. They can be regarded as paradigms indicating the predominant points of view which exist within the causal models of administrative decision programs. Five paradigms can be distinguished as being the perceived contingencies between

1. economic circumstances and residential situations
  2. economic circumstances and family or household composition
  3. employment and income
- or the focus on:
4. education or
  5. health circumstances
- without simultaneous consideration of other issues, as, for example, social circumstances.

Figure 3



<sup>14</sup> The similarity threshold chosen is 0.64 (this is equivalent to a  $\phi$ -value of 0.28). On this niveau the clusters maximally have two elements in common so that the solution is independent of additional allowed overlaps.

These distinct points of view in the representation of everyday life can be characterized as the quasi-reality of formal organizations: what is represented is not reality itself, but only a certain perspective of reality. This constructed reality is based upon concepts of that which „belongs together“ and involves masking certain other contingencies that are effectively treated as irrelevant to the purposes of the particular organization<sup>15</sup>.

### *2.3. The Optics of the Representation of Primary Environments*

Causal models within administrative decision programs can be described as a set of partially overlapping contingencies between variables measuring life sectors of their target populations as shown in Figure 1 B above. So far we have not taken into account the representation of social environments, i. e. what relationships are regarded to be relevant to the execution of administrative decision programs. For this purpose we analysed the 92 forms in terms of the different categories of persons involved. As expected, relations with colleagues or friends were almost never regarded as relevant to administrative decision programs. Only legally defined relations – as kin, household or official representatives – were found, kin and household being the most important. Accordingly we coded the 92 administrative forms in terms of 29 different legal categories of persons (see *Appendix 3*).

Both kinship and household membership are statuses that are defined legally by the administrative system. They represent two general and different philosophies toward the social atoms of concern for administrative purposes. There is, first, a genealogical perspective for identifying the descent of a person and the legally defined obligation to support relatives if necessary („Unterhaltsverpflichtung“). This orientation is important for those administrative decision programs which deal with the legal rights and duties of parents to their children. This philosophy is mainly associated with the public order administration. The second general philosophy in regard to social networks that is relevant to administrative programs is the household orientation. Households are often seen as the smallest economic unit. This concept is often independent of kinship patterns and should be expected as relevant to the housing administration. The welfare administration combines both.

Proceeding from the hypothesis that representation of egocentric networks is based on two independent dimensions – kinship and co-residence – we have carried out a multidimensional scaling analysis<sup>16</sup>. For this, a measure of similarity between

<sup>15</sup> The analysis of three population surveys showed that selectivity – as it arose in the analysis of the administrative records – is not perceived by the population. There is an almost stereotyped reduction of the diversity of the representational behaviour of formal organizations. See: Bick, Wolfgang, Müller, Paul J., *Die Buchführung der Verwaltungen als sozialwissenschaftliche Datenbasis*, in: Müller, Prozeß-produzierte Daten.

<sup>16</sup> For the literature on multidimensional scaling see the bibliography: Bick, W., et al., *Multidimensional Scaling – Theory and Applications in the Social Sciences*, Cologne 1977.

the 92 administrative tasks for all 29 categories of persons was defined: The similarity between any two tasks was measured by means of the matching coefficient<sup>17</sup>. The dissimilarity matrix was scaled using Smallest Space Analysis<sup>18</sup>. In this non-metric procedure, all 92 administrative tasks are depicted in a space of lowest dimensionality, so that the most similar tasks are distinguished by spatial proximity.

The two-dimensional solution is suitable for the data; the solution has a stress of only 0.06<sup>19</sup>. This two-dimensional solution is presented in *Figure 4*. The full exploitation of two dimensions is immediately evident.

To verify that the underlying dimensions are those of kinship and co-residence we have fitted external property vectors in the two-dimensional space in such a way as to maximize the correlation between the projections of the 92 points onto the fitted vector and the original property scale<sup>20</sup>. The following were defined as external property vectors: (1) the absolute number of represented categories of persons for one task (G), (2) the absolute number of persons in the household (GH), and (3) the absolute number of persons outside the household (GNH).

These three vectors — which could be fitted into this space very satisfactorily<sup>21</sup> — are depicted in *Figure 4*. The GH and GNH axes are almost orthogonal, and can thus be interpreted as the underlying dimensions<sup>22</sup>. The GH axis indicates the dimension „Living-together“, the GNH axis indicates the dimension „Genealogy“.

<sup>17</sup> The „matching coefficient“ D is defined as

$$D = \frac{a + d}{a + b + c + d}$$

(for the notation cf. footnote 4, for the coefficient see: Sokal, R. R., Michener, C. D., A Statistical Method for Evaluation Systematic Relationships. The University of Kansas Science Bulletin, Vol. 38 (1958), pp. 1409–1438).

<sup>18</sup> The program MINISSA-I of Guttman and Lingoës was used (see: Lingoës, James C., The Guttman-Lingoës Nonmetric Program Series, Ann Arbor 1973, pp. 39–79).

This nonmetric multidimensional scaling procedure spaces the points into a n-dimensional space of lowest dimensionality with the constraint that the similarities within the raw matrix are monotonically related to the distances in the final configuration of the solution, i. e. the rank order of the similarities is only minimally distorted.

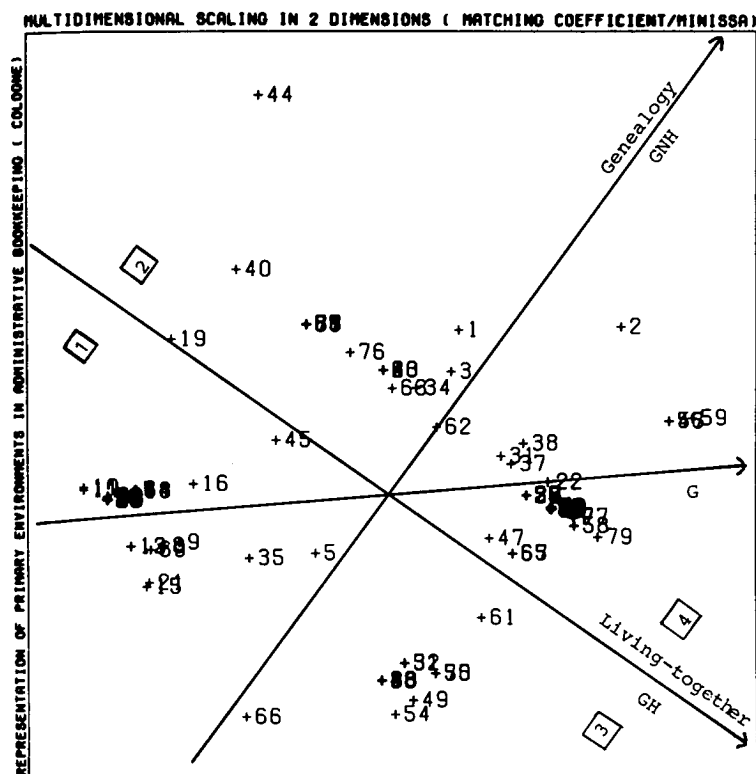
<sup>19</sup> The Kruskal's stress value is the one under condition of weak-monotonicity (Guttman-Lingoës' coefficient of alienation = 0.07). If one further reduces the dimensionality of the solution to an one-dimensional configuration the Kruskal's stress value (weak-monotonicity) is unacceptably high — 0.54 (Guttman-Lingoës' coefficient of alienation = 0.70).

<sup>20</sup> For this we used the program PROFIT from J. D. Carroll and J.-J. Chang out of the Edinburgh-Cardiff-MDS-Program Package.

For the procedure see: Carroll, J. Douglas, Models and Algorithms for Multidimensional Scaling, Conjoint Measurement, and Related Techniques, in: Green, Paul E. and Wind, Yoram (eds.) Multiattribute Decisions in Marketing: A Measurement Approach, New York 1973, pp. 299–371, see pp. 360–363.

<sup>21</sup> The product-moment-correlations between the projections onto the fitted vectors and the different property scales have the following values: for G: 0.99, GH: 0.99, GNH: 0.95. The

Figure 4



GNH = absolute number of represented categories of persons outside of the household  
 G = absolute number of represented categories of persons  
 GH = absolute number of represented categories of persons in the household  
 See Appendix 1 for the numbers of the administrative tasks.

Using these two axes, four quadrants can be distinguished:

- 1: Low representation of the primary environments of the client
- 2: Extensive representation of the genealogical relationships and low representation of the household community

angles of the axes are calculated for the case of a standardized space and drawn in the non-normalized space in figure 3.

<sup>22</sup> The final MDS-configuration is based on a distance model and therefore invariant to rotations. This allows to rotate the axes so that they correspond to the usual presentation of x-y coordinates.

3: Extensive representation of the household community and low representation of the genealogical relationships

4: Extensive representation of the dimensions „Living-together“ and „Genealogy“. The G axis, which indicates the number of persons represented runs through the quadrants 1 and 4. It runs from administrative tasks, which present the clients only as isolated persons, to tasks which represent the inclusion of the client in the social contexts of the household, the family and kinship.

The multidimensional scaling shows that the representation of categories of persons is based on the dimensions „Living-together“ and „Genealogy“, but it does not yet support the conclusion that different and distinct „optics“ exist for the representation of categories of persons.

As stated above the concepts of kinship as well as of household are legally defined social networks which try to reflect different aspects of reality. We assume that there are only limited possibilities of combining these two legal concepts without giving up their meaning. For administrative purpose the household cannot be divided into subsets whereas kinship patterns can only be divided into a few generations. Thus we expected that administrations use only a limited set of pre-defined optics this hypothesis cluster analytical procedures were applied.

We analysed the 92 x 29 (administrative tasks x categories of persons) matrix by means of entropy analysis<sup>23</sup>. This minimum variance procedure, based on the information measure, ensures detection of those groups of tasks which are inherently homogeneous, and which constitute different optics of the representation of primary environments.

Figure 5 presents the dendrogram of the entropy analysis, and the existence of natural classes, that is, optics which differ greatly for the individual groups of tasks<sup>24</sup>. The analysis of the administrative tasks included in these different clusters

<sup>23</sup> Following Vogel the entropy analysis is one of the best procedures for analysing binary data (Vogel, F., Probleme und Verfahren, pp. 109 passim, pp. 350).

The entropy analysis is based on information measures and calculates the entropy of a given group K as follows:

$$H_T(K) = m n_K \log n_K - \sum_{i=1}^m (n_{i1} \log n_{i1} + n_{i2} \log n_{i2})$$

( $n_K$ : number of elements of group K, m: number of attributes,  $n_{i1}$  resp.  $n_{i2}$  number of 0-resp. 1-values of attribute i). The procedure then fuses the two groups K and L provided that the marginal increase of the total entropy

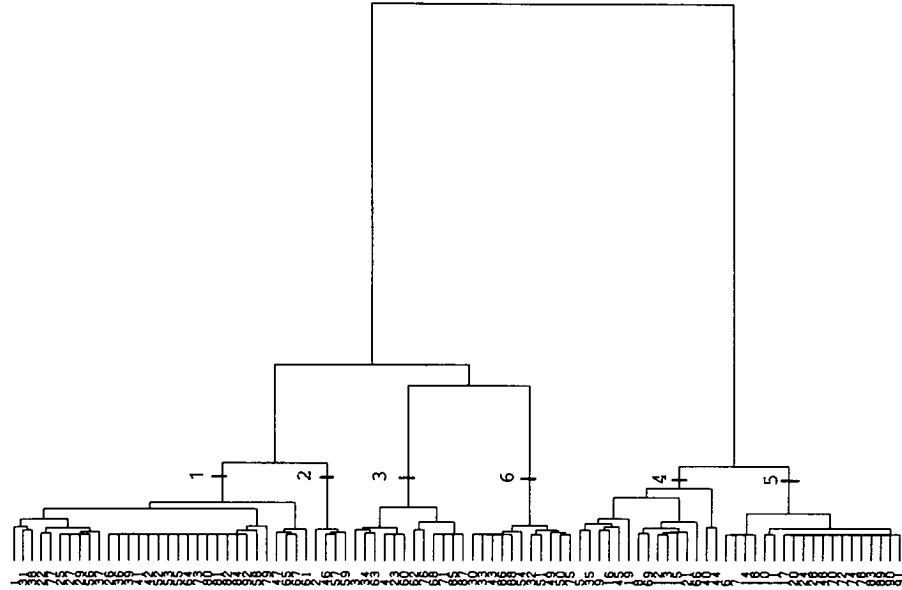
$$\Delta H_T(KUL) = H_T(KUL) - H_T(K) - H_T(L)$$

is smallest.

<sup>24</sup> These results were confirmed by applying other cluster procedures, e. g. the CONCOR-algorithm, which tries to find the so-called zero-blocks (blockmodelling). The used program was ABBW written by Clyde Mitchell.

For a description of the algorithm see: Breiger, Ronald L., et al., An Algorithm for Clustering

Figure 5



REPRESENTATION OF PRIMARY ENVIRONMENTS IN ADMINISTRATIVE BOOKKEEPING ( COLOGNE )

showed that only six different optics for representation of the social environment of clients were in use<sup>25</sup> :

- |                                     |                      |
|-------------------------------------|----------------------|
| 1 Household and family              | 4 Client and parents |
| 2 Household and relatives           | 5 Client             |
| 3 Stem family (client and children) | 6 Household          |

In *Figure 6* we have entered these optics in the final configuration of the multidimensional scaling analysis. The lines which have been drawn in define the administrative tasks with like optics.

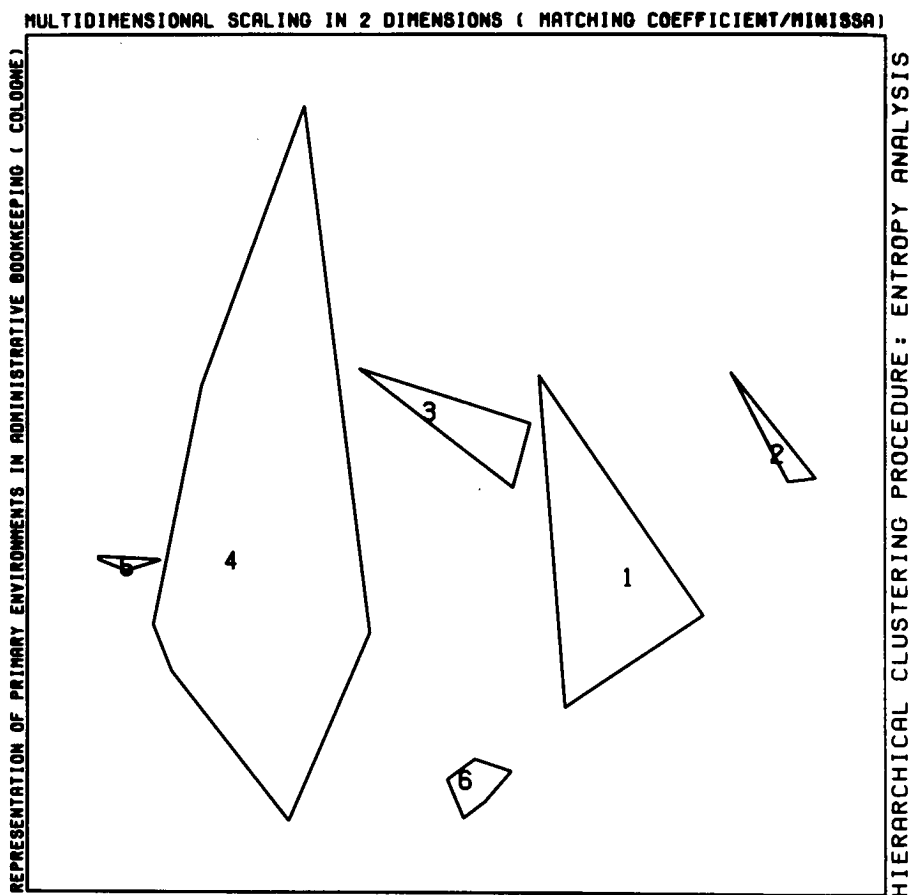
The various optics in the representation of egocentric networks dominate — as expected — in different administrative domains: the public order administration

Relational Data with Applications to Social Network Analysis and Comparison with Multidimensional Scaling, in: *Journal of Mathematical Psychology*, 12 (1975), pp. 328–383.  
When applying blockmodelling to rectangular data matrices, that is distinguishing between objects and stimuli, the concept of structural equivalence has to be modified as one of similar positions of objects vis-a-vis a set of stimuli.

<sup>25</sup> For the interpretation of these clusters the procedure RESULT was used. This procedure calculates the distributions of the variables for each cluster.

Cluster 1 includes 31 administrative tasks, Cluster 2: 4, 3: 12, 4: 15, 5: 19, 6: 11.

Figure 6



represents almost exclusively the client and his genealogy; the representation of the household and of the family typifies the welfare administration, with simultaneous use of other optics. The health insurance and the labour office are not characterized by any dominant perspectives: in these administrations, all six optics are used.

The different optics constitute concepts of the client's social environment which it is necessary to represent. They clearly do not form a continuum; their distinctiveness makes it obvious that social contexts are depicted only by means of a small set of „inter-changeable lenses“.

Our analyses have shown that there are only a few schemes used for representing primary environments of an individual in a legally meaningful way. The points of view administrations have towards these environments are therefore restrictive as

far as alternative ways of perceiving the structure of interpersonal networks are concerned. In contrast to the very small set of optics available to perceive egocentric networks, administrative causal models comprise a wide range of unique combinations of life sectors.

Although many decision programs draw upon a wide range of life sectors, they seldom share causal models totally. However, some overlap between life sectors could be found. The administrative system cannot be described as consisting of groups with the same implicit theories which are the result of belonging to the same administrative division. The system is instead a conglomerate of administrations having many individual points of view.

When administrative bookkeeping is considered in terms of the possibility of using it for sociological inquiry, the researcher is confronted with a high degree of heterogeneity in the variables represented, but also with a great opportunity of finding similar coding of primary environments in the administrative bookkeeping systems of different offices.

#### *2.4. Representing Everyday Life by Asking for Pre-Processed Information*

Because of the unique causal models within their decision programs, administrations are highly dependent on their environment for getting necessary information. Furthermore, administrative decision programs are nested in such a way that multiple client statuses must be taken into account. This leads to the necessity of linking administrative decision programs through direct communication or other forms of interorganizational linkages. But we expected that direct interorganizational exchange of information is not a sufficient way to meet the demand for information as well as for verification of information. Because of the uniqueness of the information needed, too many other organizations would have to be approached. On the other hand the risk of an unsuccessful query is high if organizations do not know all the client statuses individuals can occupy in a system of functionally specialized public administrations.

Therefore bureaucracies must ask their clients about these matters. Bureaucracies ask their clients mainly for information which has been processed in interactions between other bureaucracies and the population. The clients are not witnesses of their situation, but the organizations are. As was demonstrated in the analysis of the optics used for representing everyday life by administrative bookkeeping systems, the same reliance on pre-defined concepts can be expected as a characteristic of bureaucratic information processing<sup>26</sup>. Facts that cannot easily be trans-

<sup>26</sup> See e. g. Zimmerman, Don H., *Record Keeping and the Intake Process in a Public Welfare Agency*, in: Wheeler, S. (ed.), *On Record: Files and Dossiers in American Life*, New York 1969, pp. 319–354.



formed into facts of bureaucracies, increase uncertainty regarding the environment. Within application forms many questions therefore ask for information in pre-processed categories.

*Table 1:*

	the public order administration	the social welfare administration	other organizations	any of these
percentage of application forms that ask for links to . . .	57 %	57 %	59 %	96 %

Total: 92

As shown in *Table 1*, almost all of the 92 application forms ask for relationships to other offices. But the figures given in *Table 1* are very conservative estimates of administrative cross-validation. The 92 application forms analysed require additional documentary evidence to be provided by clients through presentation of personal documents of one kind or another. Application for rental assistance, for example, requires submission of a maximum of 14 different kinds of documentary evidence. This enormous number of appendices is not typical of this application alone, but is also characteristic of almost all application forms in the sample. Operationalizing causal models of administrations is done by using the categories of other administrations. The collection of new information or the validation of information are often done by interchanging information among organizations utilizing the individual applicants as the carriers of personal documents. This procedure implies, of course, that the administration is dependent on the cooperation of its environment to carry out this coordinating task as well as on the ability of the applicants to effectively handle such a system of symbiotic relations between clients and public bureaucracies. Unique causal models in a system of interlocked decision programs, which are operationalized by asking clients about the public statuses conferred on them by other administrations, lead to a representation of everyday life that mainly mirrors pre-processed reality. The representation of everyday life very seldom transcends those areas of reality which are not already conceived of in administrative terms. Therefore this system can be regarded as being highly introverted.

### 3. The Selective Penetration of the Population Subsystem by Urban Bureaucracies

So far, our analyses have been comparative regarding the optics as well as the causal models used to represent everyday life in the bookkeeping systems of urban bureaucracies. We now wish to demonstrate the extent to which public bureaucracies represent everyday life selectively and to what extent the selectivities of various bureaucracies cannot be overcome by interlocking bookkeeping systems. We shall identify the „holes“ in the representation of everyday life within bookkeeping systems which lead to a partial blindness of the administrations when perceived as a system.

Application forms ask for those life sectors which are relevant to the administrative concern. To find out the holes within administrative bookkeeping — that is, those life sectors and egocentric networks that are blocked out by the administrative system — we again used the standardized application forms as data sources.

In *Figure 7* we have drawn the selective screens used in representing everyday life of a client during his life course. A sample of 15 administrative forms, which can be regarded as the set of forms almost everyone has to fill out during his life, were coded according to 15 life sectors. This figure shows the effects of the unique causal models which are implicit in administrative decision programs. Those life sectors which are not covered by administrative forms will be termed the holes in the administrative bookkeeping systems. But the selectivities inherent in the single administrative programs could well disappear when the cumulative effects of these representations over time are considered.

As shown in *Figure 8*, this is only partially true. *Figure 8* shows the representation of life sectors as well as of the most important legally defined egocentric networks over time. *Figure 8-A* depicts the representation of everyday life in the normal case, whereas *Figure 8-B* presents the results for an individual who applies for a variety of services within the social welfare administration, that is, for an individual who additionally applies for subsidy to occupational training, young people's welfare, unemployment compensation, public assistance, and rental assistance.

It is evident that the population subsystem is penetrated differentially, dependent on the need of individuals to rely on governmental services and that, as a consequence, representation of everyday life of people in lower strata or in bad economic situations is much more complete. But as only a small proportion of the population comes into contact with a wide range of administrative services, the representation of the normal cases shows the deficiencies of administrative bookkeeping much clearer. Administrative bookkeeping very often does not reflect the housing conditions or the economic and health conditions of those people who are members of the applicant's household. Administrations within urban bureaucracies are normally innocent of information about these life sectors or states of affairs of primary

environments. Due to their partial blindness, changes within these holes can hardly be recognized, although they may heavily influence the target variables of some administrations.

Figure 7

Attributes	Applications	notification of birth	registration	school-health card	military registration	wedding notice	passport	issuing a social security number	health insurance registration	issuing a state wage tax card	application for reimbursement of taxes	driving license	car registration	subsidy to children	application for old-age pension	notification of death
personal data																
family and household																
residence																
employment																
education/qualification																
economic situation: income																
economic situation: expenditures																
health																
memberships																
military																
special population groups																
links to the public order administration																
links to the social welfare administration																
links to other organizations																

Figure 8

(A)

Normal Client

(B)

Multi-Problem Client

	living in the household					living outside the household			living in the household				living outside the household			
	applicant	spouse	children > 18 years	children < 18 years	parents	children > 18 years	children < 18 years	parents	applicant	spouse	children > 18 years	children < 18 years	parents	children > 18 years	children < 18 years	parents
personal data																
family and household																
residence																
employment																
education/qualification																
economic situation: income																
economic situation: expenditures																
health																
memberships																
military																
special populations																
links to the public order administration																
links to the social welfare administration																
links to other organizations																

#### 4. The Selective Use of the Network of Institutions for Administrative Bookkeeping

It was assumed that the selectivity found to be characteristic of information-gathering would also distinguish the use of the network of institutions for exchange of information, e. g. for the enrichment and reciprocal confirmation of information or files. It was possible to confirm this hypothesis by means of a survey of administrative clerks which we carried out in the Cologne area<sup>27</sup>: primarily only that information is exchanged between administrative units which already played a role in the gathering of information from clients. There is little in the way of exchange for the sake of enrichment of information.

It follows that vulnerability to error in the bookkeeping of a network of institutions exists when the circle of parties who engage in interaction constitutes only a selective segment of the total number possible and when in the network of institutions there are sectors or zones of dense interaction which are characterized by a high degree of internal communication and low inter-sectoral communication. Given such structural constraints, errors which arise in the gathering of information by one unit can be passed on to the others; that is, these errors are not perceived, and thus are reproduced.

In order to check this preliminary hypothesis we carried out a secondary analysis of a study which included all written communication of 64 city administration offices in Nuremberg<sup>28</sup>. In 1970 about 5000 employees of the Nuremberg city administration enumerated all their contacts with one another. This study was carried out by the organization office of Nuremberg with the objective of providing assistance in efforts to economize. In our secondary analysis, this study is consulted for the analysis of interorganizational information behavior. The communication matrix which resulted from the above mentioned study was analysed using multidimensional scaling. The flow of information structures the administrative offices into a network with sectoral differentiation.

<sup>27</sup> In the standardized survey within administrative units of the City of Cologne 96 clerks were interviewed for those administrative tasks which — similar to the selection of the 92 administrative tasks — are client-oriented and in which standardized forms are used. The survey was started in autumn 1976 and was finished in spring 1977. For the conception see: Bick, W., and Müller, P. J., *Die informationelle Abbildung der Klienten in formalen Organisationen — Konzeption für eine empirische Untersuchung*, mimeo, Köln 1976.

<sup>28</sup> The communication frequencies were collected in 1970 by the organization office of the City of Nuremberg: For a fortnight about 5000 administrative officers of the City of Nuremberg counted their communication distinguishing written, telephone, and face-to-face contacts in and outside of the department. See: *Kommunikationsanalyse 1970, Stadt Nürnberg — Organisationsamt, Untersuchungsbericht und Beilage 1: Tabellensammlung und Beilage 2: Graphische Darstellungen*, Dezember 1970.

*Figure 9* presents the results of the Smallest Space Analysis<sup>29</sup>. Sectoral differentiation was roughly sketched in: grouped around the Administration of Resources are the sectors labelled Building, Public Order, Social Welfare and Health, Schools and Culture. These sectors are not identical with the formal organization of the Nuremberg city administration, however. This is one indication that the principles of „integrative centrality“ and of „sector differentiation“<sup>30</sup> cannot be derived from the formal structure of city administration.

Multidimensional scaling confirmed the principles of „integrative centrality“ and „sector differentiation“. But the validity of these principles is limited by the fact that communication densities were defined only bilaterally, and the final configuration can be interpreted only globally in the sense of its „Gestalt“. The hypothesis of existence of zones of dense interactions which can be distinguished from one another according to principles of external isolation can be approached only by analysing the structure using network analysis. For this we used the concept of cliques and constructed fields of dense interactions<sup>31</sup>. Two domains of interacting offices on the community level could be identified: people processing and object processing organizations. People processing organizations are defined as those which classify, confer public statuses and dispose clients<sup>32</sup>. Object processing organizations are mainly concerned with handling buildings and land. Within these two domains administrations exchange information and often solve their problems in coordinated efforts. Both domains are sketched into *Figure 9*. It becomes evident that sectors do not correspond to domains of interacting offices. The people processing domain cross-cuts sector boundaries; the Schools and Culture sector does not build up a domain.

<sup>29</sup> The program MINICPA by Roskam was used (ISF = 1, METHOD = -1, i. e. hard squeeze 1, monotone regression). See Roskam, Edward E., *Nonmetric Data Analysis, General Methodology and Technique with Brief Descriptions of Mini-Programs*, Report 75-MA-13, University of Nijmegen, Department of Psychology.

This non-metric multidimensional scaling program works as follows: For each unit *i* the other units are placed in a space of lowest dimensionality so that those administrative units receiving most information from *i* (row solution) resp. sending most information to *j* (column solution) are close to unit *i*.

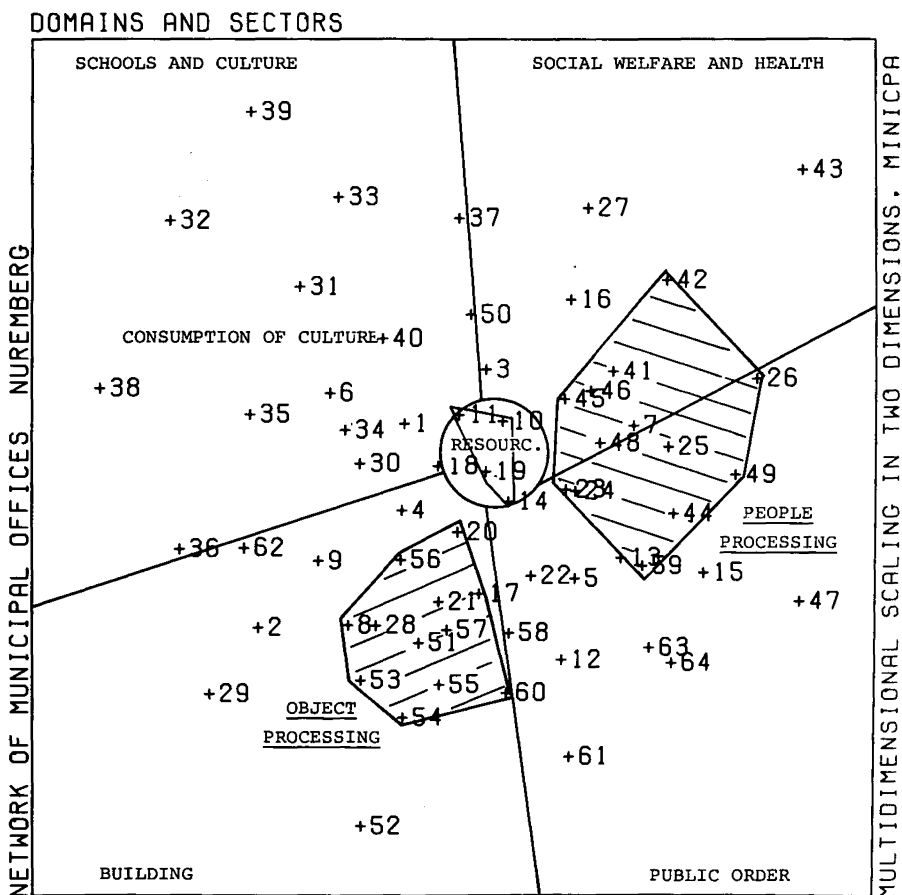
The solution given is based on a row solution. Its coefficient of alienation of 0.198 is acceptable taking into account the large number (64) of units. The column solution leads to a similar result.

<sup>30</sup> Laumann and Pappi obtain similar results when analysing the differentiation of a network of elites in a local community. See: Laumann, Edward O., and Pappi, Franz U., *Networks of Collective Action, A Perspective on Community Influence Systems*, New York 1976.

<sup>31</sup> We applied the clique detecting program NCLIQUE. See: Felling, A. J. A., *A Graph-Theoretical Approach to the Structure of Local Elites*, in: *Zeitschrift für Soziologie*, Vol. 4, No. 4, pp. 221–233. For the procedure applied to construct the domains see: Bick, Wolfgang, and Müller, Paul J., *Stable Patterns within a Network of Urban Bureaucracies – Domains or Positions?*, paper presented at the 1978 American Sociological Association Meeting, San Francisco 1978.

<sup>32</sup> See: Hasenfeld, Yeheskel, *People Processing Organizations: An Exchange Approach*, in: *American Sociological Review*, Vol. 37 (1972), pp. 256–263.

Figure 9

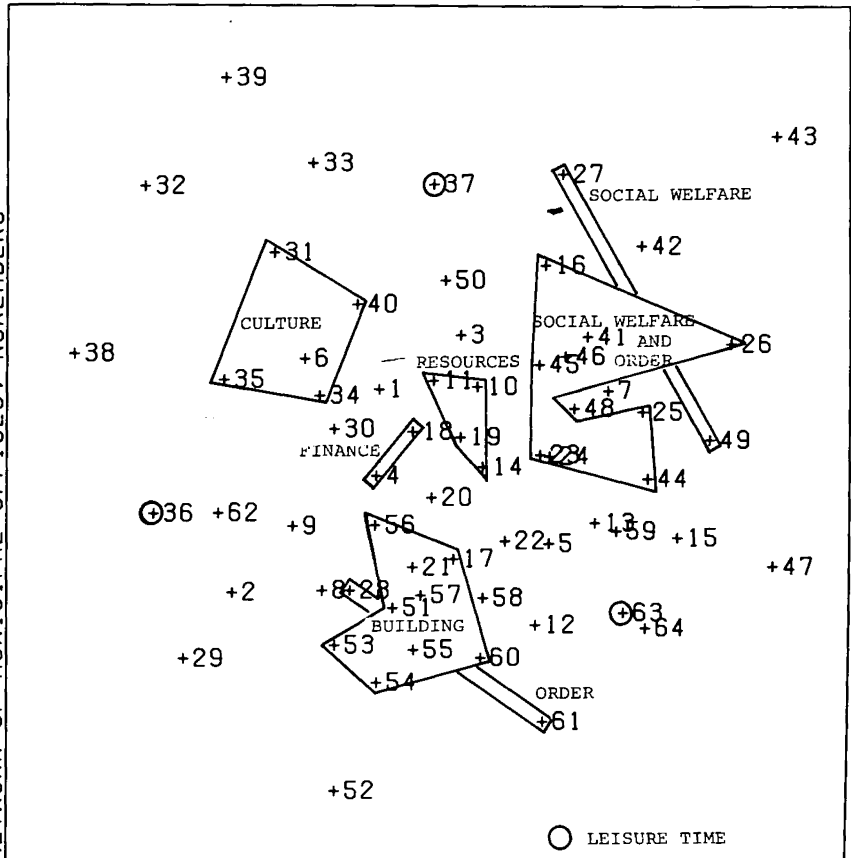


Stress  $\hat{d}_{hat} = 0.189$

See appendix 4 for the identification of the 64 offices.

Vulnerability of a network of information exchange not only depends on direct communication between offices. Offices may interact with similar sets of other offices without necessarily interacting with one another. The set of organizations approached by one organization is the organizational environment of an office. This environment determines the kind of information that can be acquired from other offices thus leading to structural constraints on learning of those aspects of reality that are processed within organizations not belonging to one's own organization set.

NETWORK OF MUNICIPAL OFFICES, NUREMBERG



<sup>33</sup> The conception of the „organization set“ was developed by Evan as an analogon to Merton's role set. See: Evan, William M., *The Organization Set: Toward a Theory of Interorganizational Relations*, in: Thompson, J. D. (ed.), *Approaches to Organizational Design*, Pittsburgh 1966.



circles"<sup>34</sup>. This approach made it possible for us to ascertain the cohesive groups of institutions with high degrees of overlapping interaction partners. To do so we used Johnson's hierarchical clustering procedure<sup>35</sup>. The following groups could then be identified: Culture, Leisure Time, Social, Social/Order, Order, Building, Finance, Resources.

In Figure 10 we have entered these groups into the two-dimensional solution of the Smallest Space Analysis which was presented in Figure 9. These local groupings constitute those groups of offices which are characterized by almost the same organization set. They have the same selectivity pattern in the use of institutions for the enrichment or confirmation of information within their systems of bookkeeping.

The vulnerability of the bookkeeping of the network of institutions on the community level can be traced to the fact that, on the one hand, administrative tasks are solved within limited sets of interacting offices, and that, on the other, elements of the networks are supported by the same „correspondents' network“. These factual information barriers set the limits to information which can be found in the files of the respective offices. The barriers constitute institutional limits to administrative bookkeeping. They exist as additional limitations, besides the informational limitations which could be seen through the analysis of the information-gathering behavior reported above. The possibilities of institutions exchanging information are not only limited in terms of selectivity in representing individual attributes or social contexts, but also in terms of selectivity in the use of the administrative environment<sup>36</sup>.

<sup>34</sup> See: Simmel, Georg, *Soziologie. Untersuchungen über die Formen der Vergesellschaftung* (1908), 5th ed., Berlin 1968, pp. 305–344. For the calculation of the similarities the following formula was used:

$$n_{ij}^P = \frac{(n_{C_i} \cap n_{C_j}) - 1}{(n_{C_i} \cup n_{C_j}) - 1}$$

( $n_{C_i}$  consists of all elements  $j$  with a path length of  $n$  or less between  $i$  and  $j$ ). For our analyses we have used the subroutine SIMMEL 2 in SOCK ( $n = 1$ ).

For the similarity measure see: Alba, Richard D., and Kadushin, Charles, *The Intersection of Social Circles, A New Measure of Social Proximity in Networks*, in: *Sociological Methods and Research*, 5 (1976), pp. 77–102.

<sup>35</sup> For the application of Johnson's hierarchical clustering procedure we have used the subprogram DIAMETER in SOCK. This procedure is similar to the complete linkage approach described in footnote 9.

The labels used again have no resemblance to the definitions of divisions within the city administration.

<sup>36</sup> Concerning the interdependence of informational and institutional constraints Jensen obtains similar results: „Most sender and receiver components in the administrative system are related to a certain division . . . with which it communicates in particular, but . . . this kind of specialization is not followed by a similar specialization with respect to the kind of information sent or received by these components“. (Jensen, Mogens Brabrand, *Informatics and the Cen-*

## 5. The Quality of Administrative Bookkeeping from the Point of View of Administrative Personnel

In the survey of administrative officers in Cologne, we wanted to know how the filing personnel themselves assessed the reliability of their records. In order to relate the answers obtained in this survey to the structural properties of the representations, we combined the interview data with the data from the analysis of the 92 administrative tasks<sup>37</sup>.

This record-linkage made it possible to develop an external definition of the problems of representation which are faced by the different administrative tasks; that is, a definition independent of the subjective assessments of the clerks interviewed. For this purpose we calculated for each of the 92 administrative tasks, the absolute number of the attribute groups represented for nine categories of persons<sup>38</sup>. These profiles — in the following termed „representation screens“ — indicate the intensities of administrative interest in information about the applicants and their primary environments. We analysed the resulting data matrix using Ward's procedure and by means of iterative relocation of the resulting clusters<sup>39</sup>.

tralization Issue. A Danish Case, Institute of Public Administration, University of Copenhagen, Denmark, Report to the 16th Congress of the International Institute of Administrative Sciences, Mexico, July 1974, p. 77).

<sup>37</sup> Because of the differentiation of administrative tasks in intake and processing more than one interview was made for certain tasks. Those interviews with personnel that is exclusively concerned with client processing were not included in the following analyses. This leads to a reduction from 96 to 84 interviews. Furtheron for some administrative tasks we did not administer an interview. The assignment of screens of representation to individual interviews was done on the basis of all 92 administrative tasks.

<sup>38</sup> The following categories of persons were distinguished: client, spouse, children, parents, grandparents, grandchildren, siblings, other relatives, strangers.

<sup>39</sup> The Ward procedure first calculates the error sum of squares for a given partitioning:

$$E_K = \sum_{k=1}^{n_K} \sum_{i=1}^m (x_{ikK} - \bar{x}_{iK})^2$$

( $n_K$ : number of elements of group K,  $m$ : number of attributes)

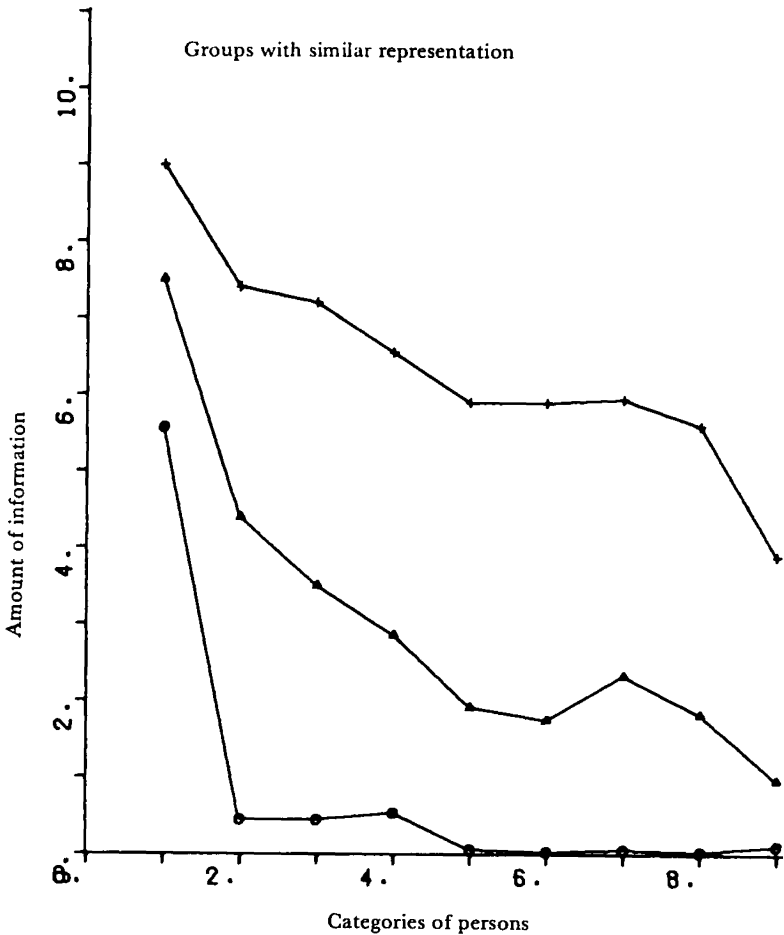
The procedure then proceeds to a fusion of the groups K and L which leads to the minimum increase of the total error sum of squares

$$\Delta E_{K \cup L} = E_{K \cup L} - E_K - E_L$$

This was computed by using subroutine HIERARCHY, coefficient 6. For improvement of the found clusters we used the iterative relocation procedure RELOCATE in CLUSTAN 1 C. In this procedure all elements of cluster X are assigned to a different cluster Y if the distance to the centroid of Y is less than the distance to the centroid of X.

The administrative tasks with similar representation screens were assigned to three classes. *Figure 11* illustrates the differing intensity, in regard to the representational range: Screen 1 with the least, Screen 2 with medium and Screen 3 with the greatest representational intensity.

Figure 11



- Screen 1: 36 administrative tasks
- ▲ Screen 2: 36 administrative tasks
- + Screen 3: 20 administrative tasks

We have established various indicators for the assessments of representational quality by the filing personnel:

1. The clerk's assessment of the client's full exploitation of all self-portrayal possibilities (competence)<sup>40</sup>;
2. The clerk's assessment of the extent of the client's difficulty in assigning himself to the categories prescribed by the administration (particularism)<sup>41</sup>;
3. Assessment of whether files compiled by other clerks are on their own sufficiently expressive to give a correct picture of a client not personally known to the file reader (file image)<sup>42</sup>;
4. The clerk's assessment of whether or not there is any possibility of checking the veracity of the data given by the client (helplessness)<sup>43</sup>.

In addition, two dimensions of administrative information behavior were measured:

1. The distrust of given information, expressed in the clerk's attempt to verify clearly plausible data<sup>44</sup>;
2. The frequency with which the clerk calls attention to possibly relevant data which has not been given<sup>45</sup>.

<sup>40</sup> This indicator was measured by means of the following question: Now a question about the initial applications which you receive completely filled in by the applicant. Do you have the impression that the applicant has included everything which could be of use to him?

Yes: 61 % No: 18 % N/A: 19 % No answer: 2 %.

<sup>41</sup> Have you noticed that the applicants call attention to facts which play no role in the processing of their application? We are referring especially to statements like „You have also helped my relatives“, or, for example, „I still have to pay for my car“ or something similar, which is included in the conversation.

Yes: 73 % No: 27 %.

<sup>42</sup> Thinking back to situations in which you have processed files which were compiled by other clerks, are the applicant's circumstances always clear in such files? Can you always get a correct picture of the applicant or, do you most often prefer to refer back to the applicant? Please give your answer using the card. Which answer is most applicable to you?

File always provides sufficient information	Yes: 39 %	No: 61 %
---	-----------	----------

File usually provides sufficient information	Yes: 43 %	No: 57 %
--	-----------	----------

Usually refer back to applicant	Yes: 16 %	No: 84 %
---------------------------------	-----------	----------

Other	Yes: 7 %	No: 93 %
-------	----------	----------

<sup>43</sup> In the application for . . . there are questions to which the citizen answers no, and thus needs not present any certification. If, for example, someone says that he is not receiving a service from an office, he won't present any certification from that particular office. What can you do in such cases: do you always have to believe the citizen, or are there also possibilities of verifying such answers elsewhere?

Believe: 37 % Verify elsewhere: 47 % Other: 16 %

<sup>44</sup> In the applications there is a lot of data which at first glance appears to be quite clear. Do you usually check once more whether or not erroneous data might have been given?

Yes: 45 % No: 55 %

<sup>45</sup> It is not always clear to the applicant just what all can be included in an application. In such

The more comprehensive the representation screen, the less adequate is the information provided by the file (screen/ file image,  $\gamma = -0.40$ )<sup>46</sup> – The more the clerks „ask“, the less they feel they „learn“.

The filing clerks thus feel that wide-mesh representation screens result in less reliable and adequate representations. This apparently paradoxical finding is resolved by considering that with increasingly comprehensive representations, the assumed competence of self-portrayal declines (screen/competence,  $\gamma = -0.37$ ), and the difficulty of assignment to the administrative categories increases (screen/particularism,  $\gamma = 0.46$ ).

These difficulties of assignment are one indication of tensions which occur in the representation of everyday life because of the selective screens of formal organizations. Correspondingly, the clerks feel less confident with wide-mesh representation screens than with other information-gathering screens. This phenomenon of „asking more but learning less“ will be referred to hereafter as „decreasing marginal use of additional information“.

However, the greater uncertainty involved in administrative tasks with wide-mesh representation screens parallels the assessment of decreasing helplessness in regard to ability to alter this situation (screen/helplessness,  $\gamma = -0.43$ ): exactly those clerks who sort clients into relatively wide-mesh representation screens claim to have possibilities for verifying the validity of data at other administrative offices.

The different degrees to which clerks distrust the information given them by their clients do not vary with the representation screens ( $\gamma = 0.04$ ); on the contrary, the assistance provided by the clerks increases proportionately to the complexity of representation ( $\gamma = 0.49$ ), independent of the perceived competence of the client<sup>47</sup>.

We hypothesized that one of the distinguishing features of formal organizations would be that formally complete applications are defined as unproblematic. A consequence of this labelling as „unproblematic“ would be that the clerks would not undertake any further verification of formally complete applications. Paradoxically, only an incompletely filled-in application or questionnaire would then stimulate the clerk to gather additional facts and in this way establish the possibility of a representation of the client which would tend to be more in keeping with actual facts.

cases do you often ask the applicant about certain circumstances which he has not included in his application, but which could be important in the processing of his case?

Yes: 58 % No: 27 % N/A: 15 %

<sup>46</sup> This definition of the intensity of representation combines both – the depth and range of representation. Besides this definition we have chosen other operationalizations which only included one of the two dimensions (number of compiled categories of persons, number of compiled attributes of the client, average number of compiled attributes for all categories of persons). All results presented remained stable for these variables dichotomized at the mean.

<sup>47</sup> In general there was a correlation between the willingness to help and the perceived competence: the lower the perceived competence the higher the willingness to help ( $\phi = 0.37$ ).

And our survey did indeed discover that the clerks equate formal completeness with adequacy of representation: 77 % of all those interviewed have the impression that in the case of fully completed applications, the applicant has included all information of use to him.

However, the equation of formal completeness with adequacy of representation occurs more often in the medium- or fine-mesh representation screens ( $\gamma = 0.37$ ). Thus, the uncertainties involved in wide-mesh representation screens, as stated above, must be supplemented. Besides the phenomenon of decreasing marginal use of additional information in wide-mesh representation screens, there is the problem that errors in formally complete data in fine-mesh representation screens are not detected.

If we look at the distribution of representation screens in the administrative domains we perceive the diversity of representational deficiencies with which these fields are confronted: the public order administration — executing classical state functions —, the health insurance and the majority of the tasks of the labor office use predominantly the finest-mesh representation screens, while the welfare administration — as a representative of modern welfare and welfare planning systems — uses the most comprehensive representation screens.

In its administrative bookkeeping, the former group is confronted with the representational deficiency resulting from the failure to detect errors, and the latter with the problem of the inadequacy of its representations. This same deficiency will plague the administration of new societal problems, because increasingly comprehensive screens are demanded for the representation of clients.

## 6. Public and Private Contingencies of Everyday Life

Causal models within administrative decision programs are specific to the focus of administrative concern although nested with other causal models through interconnections to other offices. The causal models can therefore be regarded as being unique but as having overlapping points of view. In contrast, the points of view individuals have towards their social situation should be described as a web of contingencies. They are diffuse as the contingencies are not focused on a specific point of view. Confronted with the questions asked by bureaucracies, people transform their contingencies into explanations. These explanations are to some extent dependent on the variable to be explained, which leads to a variety of different causal models. Being asked for working conditions, people might not mention their housing conditions, but instead talk about their health. Asked for the ability to move as a prerequisite to taking a new job, people might also think about their current housing situation, but not about their health conditions. As bureaucracies ask for different

information they activate the web of private contingencies and are then confronted with a whole set of different private causal models, showing the difference between bureaucratic and private contingencies.

In a survey of the clients of various local bureaucracies done in Cologne in 1978, we investigated the extent to which administrative causal models and private contingencies overlap. As reported in *Table 2*, the public assistance program is confronted with a high degree of non-overlap concerning the definition of what is and what should be regarded as relevant to determining appropriate actions. The percentages given indicate that administrative decision programs are differentially evaluated in regard to the degree to which the philosophies of administrations coincide with private contingencies.

*Table 2*

	Has everything been discussed which was of interest to you?		N
	% Yes	% No	
Public assistance	51	49	147
Rental assistance	77	23	135
Unemployment compensation	65	35	159
Medical rehabilitation	65	35	48

This result however does not indicate in which ways the differences in perceptions operate. Differences between private and public contingencies could emerge in two forms:

1. The categories used in the administrative decision programs are not the categories used in everyday life. Whereas the administrative categories are very specific and only legally defined, we expect the citizens to have a much broader understanding.
2. Questions asked by bureaucracies are often regarded as being too narrowly posed. They can therefore often not be adequately answered without an explanation. Questions asked by bureaucracies concerning a specific life sector are then answered by referring to private contingencies. These private contingencies may transcend those contingencies implied by the causal models of administrations.

In an observation study of the encounter between clients and bureaucracies done in 1978, we measured the differences between private and public models of everyday life contingencies. The answers given by the clients to questions concerning a life sector were coded according to the life sectors mentioned by the client. Answers of clients were coded if they encompassed much more information than required by the civil servant or even transcended the life sectors asked for. In this way we

measured the perceived necessity for clients to give individual explanations in their answers.

In *Figure 12* we have drawn the correlations of everyday life as evoked by asking questions concerning the three most important life sectors within the public assistance program: work conditions, income and expenditures. Based on a total number of 122 screening interviews observed, work conditions were asked for in 52 %, income and expenditures in 75 % and 57 % respectively. The answers to the questions are portrayed in the middle of the figure. In 20 % of all questions concerning employment the client did not share the administrative definition. The corresponding values for income and expenditures are 23 % and 19 %, thus indicating that administrative categories are often seen as too specific from the individual point of view.

Finally, we have drawn in those contingencies which — although lying in the range of administratively acceptable contingencies — describe what necessarily belongs together according to the private causal models. The clients cannot discriminate so well between the three life sectors, but this is dependent on the focus asked for. Employment is seen in connection with its income-generating function, whereas expenditures are seen in connection with income, therefore being a net concept. Income is perceived in connection with both other dimensions.

As the function of the screening interview is to determine the eligibility to become a client of an office, the results do not necessarily indicate the degree of concordance between bureaucratic and private definitions, but instead the different weights within sets of relevant dimensions as seen by bureaucracies and their publics. One should expect that administrations are confronted with a low degree of congruence between their definitions and the clients' definitions if they use very specific and selective definitions of life sectors which are much more diffusely and inclusively conceived of in everyday life. That is, the selectivity of administrative categories is not accepted by the client population.

To determine the amount of missing congruence between administrative categories and private connotations we have analysed the observed process of jointly filling out a standardized questionnaire within the offices of welfare and youth, the local labour office, the public housing administration, and the local health insurance. We measured the frequencies with which clients were not able to answer the questions as concisely as required by the application form. The results of our analyses are reported in *Table 3*. The percentages given indicate the extent to which the civil servants in various offices are confronted with different definitions held by individual applicants regarding administrative categories. The lower the percentages given, the more people agree with or accept the administrative definitions.

Income is the most controversial concept used within the offices analysed. The local labour office is not confronted with a major lack of congruence between administrative categories and private connotations concerning unemployment compensation, whereas the public housing administration uses only a few categories within the application for rental assistance which have different meanings for the client population.



Figure 12

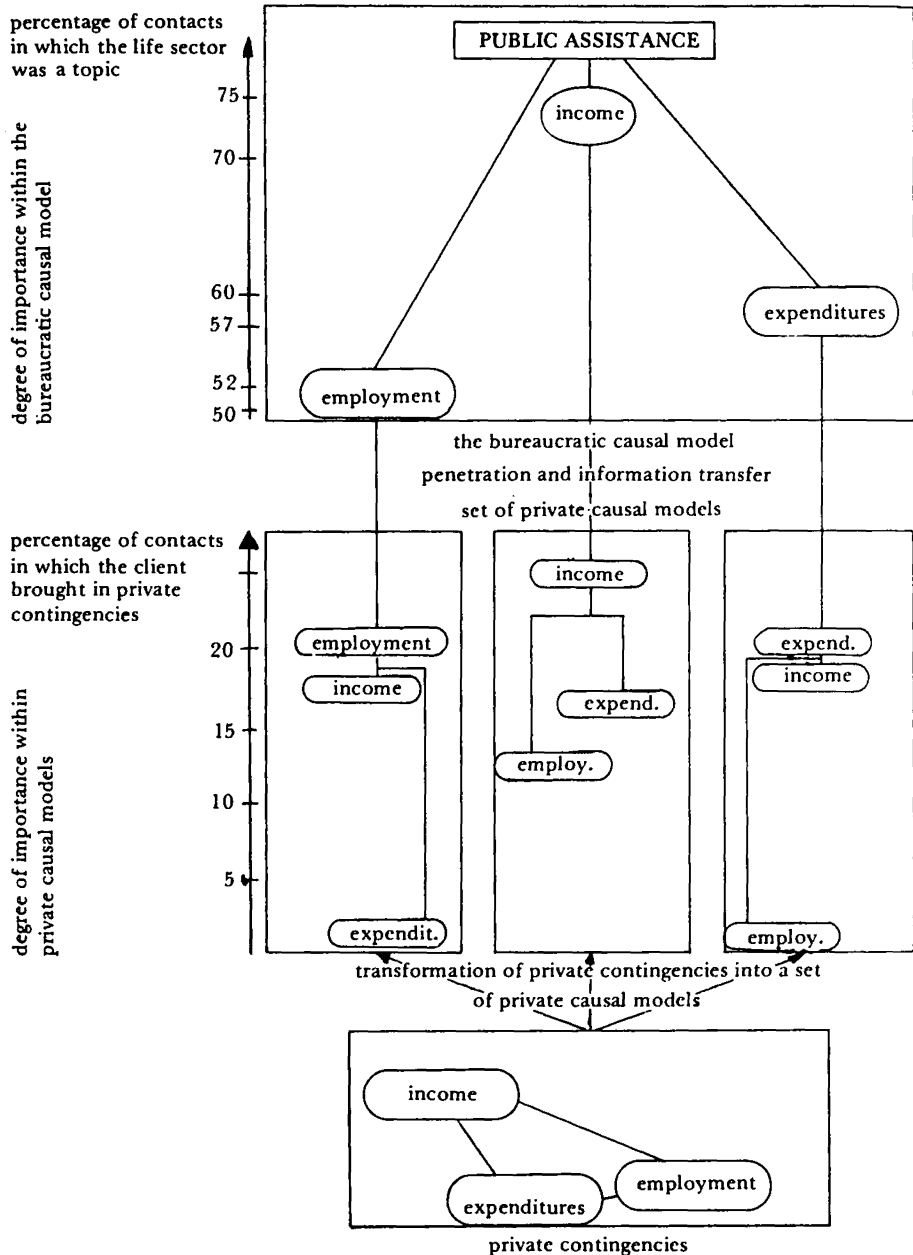


Table 3

	unemployment compensation	rental assistance	subsidy to educational training	medical rehabili- tation	low rental housing license	public assistance
education	—	—	38	—	—	—
employment	38	—	—	—	—	44
income	16	55	61	—	73	42
expenditures	—	36	—	—	—	42
health	—	—	—	63	—	—

— indicates that these life sectors were not asked for (less than 50 %).

The license for eligibility to live in low rental housing, the application for medical rehabilitation, and the application for a subsidy to educational training are all characterized by specific categories which have different meaning for the client populations.

The administrative definition of the concept of income is not shared by those applying for a low rental housing license, whereas the office granting a subsidy to educational training is additionally confronted with a problematic concept of education. Being asked for their health, clients refer to the cause as well as to consequences of their health condition. Health is mainly regarded as an intertemporal concept, which means that questions concerning health cannot be answered without explanations. In addition to that, there exists no common measurement scale to determine the degree of sickness. Therefore, local health insurance is confronted with a high degree of divergent perceptions of health within its client population.

In fact, administrations in varying degrees are confronted with divergent perspectives within their client populations. With the exception of the labour office the very specific categories used within public offices are not fully shared by the clients. This result can be seen as the consequence of bringing private connotations into the process of „skeletonizing“ people's problems.

However, the percentages given for the public assistance program indicate a seemingly paradoxical result. The office that has the broadest understanding of the concepts used is still heavily confronted with the problem of different connotations and definitions within all life sectors. This paradox can be explained as a consequence of the perception of the administrative „client model“ by clients. Clients offered an opportunity to express a wide variety of contingencies activate their own private causal models in the encounter situation and express their own connotations of administrative concepts. The less administrations are restrictive in their client models — that is, the more they are oriented to the whole biographical space of the client —, the more they are confronted with the complex web of contingencies operating in everyday life. Being still a bureaucratic organization interested in classifi-

cation of clients means that the office cannot accept all the contingencies it invites. Even if administrations are less selective, clients still disagree with the ways they are informationally processed. If administrations try to reduce the distance between private and public problem definitions by decreasing the selectivity of representation, the distance between bureaucratic models and everyday life is still apparent.

If administrations are frequently confronted with private contingencies and private connotations of administrative concepts, the classification of clients can only be done by a funnelling process. This process must translate the information given by individual applicants into administratively usable definitions. This translation may lead to a much better representation of everyday life within administrative bookkeeping when done within the encounter. In those cases, in which the „skeletonizing“ of individual contingencies into the reduced set of administrative categories is done by the clients themselves, administrations cannot even learn how biased their bookkeeping may be.

## 7. Potential Use of Process-Produced Administrative Data

We have demonstrated how the process of categorical reduction influences the perception of reality by the administrative system and the way in which bureaucracies attempt to put together their fragmented points of views. Functional specialization within urban bureaucracies, and the corresponding differentiation of causal models that are used to perceive the environment of organizations, lead to neglecting aspects of everyday life which are not regarded as being of administrative concern as well as certain interdependencies. Although urban bureaucracies form a network and exchange information about various aspects of their environment, the system of urban bureaucracies cannot internally identify its partial blindness or its misperceptions of contingencies operating in everyday life.

Up to this point we have described the structural deficiencies of administrative bookkeeping data. These conditions have not yet been investigated sufficiently to meet the need for social-scientific source criticism for a given data base. It would appear to be possible to meet such demands only within the framework of research projects which analyse concrete and specific data files.

On the other hand, the special advantages of process-produced data for sociological research<sup>48</sup> must be pointed out. Process-produced data have their own

<sup>48</sup> As to the problems of documentation, archiving, and access with the use of administrative bookkeeping data see: Müller, Paul J., Administrative Bookkeeping as a Social Science Data Base, paper delivered at the August 1976 IASSIST-meetings, mimeo.

merits. As they are reflections of the activities of formal organizations in a society they indicate influences on individual actors within society rather than the opinions of those actors.

The traces of everyday life which are provided by the bookkeeping of institutions are especially useful for the analysis of processes. The special nature of many administrative bookkeeping systems as records of change – not states but changes are reflected – corresponds to the analysis of life histories at their „turning points“. Other bookkeeping systems are distinguished by their almost continuous representation of clients. For that matter, the latter systems are suitable for the analysis of both, constancy and change in life histories.

Both of these approaches could be used as substitutes for attempts to collect data through individual recall of processes – like the passing of various positions in the life cycle – through such devices as individual interviews or diaries as used in time budget studies<sup>49</sup>. Instead, both approaches would read the changes in the mirrors provided by the bookkeeping institutions. Independent of the methodological problems involved in the ascertainment of processes by means of interviews or self-observation, there is a distinguishing characteristic of the representation of interactions or interaction sequences by administrative bookkeeping systems. The client learns only of the direct contacts to bookkeeping authorities, and is often not aware of the cross-connections<sup>50</sup> with other institutions that result from communication within the network of institutions. Given this situation, the records of the administrative system concerning the persons represented become more important than the statements of the persons about their environments.

If we are interested in the ways in which individuals are tied into a social system characterized by record-keeping institutions, the representation of individual networks within the administrative bookkeeping becomes increasingly important both for contextual analyses, and for the analysis of processes over time, using individual networks as units of analysis. For many administrative tasks the representation of just the isolated individual is inadequate. It is precisely this point which makes it possible to fall back on representation of social networks – either directly or by combining several different data bases.

The combination of administrative data makes it possible to link observations to various points in time. It also gives rise to a linking of files with egocentric networks as units, or of persons who are represented as clients in various administrative bookkeeping systems.

<sup>49</sup> See: Szalai, Alexander (ed.), *The Use of Time, Daily Activities of Urban and Suburban Populations in Twelve Countries*, The Hague 1972.

<sup>50</sup> For the problem of informational autarky see: Müller, Paul J., *Informationsflüsse und Informationshaushalte*, in: Steinmüller, Wilhelm (ed.), *Informationsrecht und Informationspolitik*, München 1976, pp. 95–109, Jensen, Mogens B., *The Use of Data Banks in Public Administration – Organisational Consequences and Political Responsibility*, in: *The First International Oslo Symposium on Data Banks and Society*, Oslo 1972, pp. 27–48.

In both of these procedures, the evolving data bases can take the form of „complex data files“. This creates data management and analyses problems which cannot be solved within the data analyses packages which up to now have concentrated on rectangular  $n \times m$  matrices.

From previous applications of process-produced data, it is obvious that there has been hardly any attempt to deal with the kind of research questions which we have described. Instead, there has been a preoccupation with analyses of process-produced data which proceeded as if the data had been gathered by isolated surveys, and not by a system of interlocking bookkeeping<sup>51</sup>. Characteristic of this orientation is the dominance of studies of selected attributes of the clientele of one or more organizations, for example the „social structure“ or social situation of the clients of welfare organizations. But such analyses of process-produced data are more seriously confronted with the problem of administratively preconceived reality in such representations than approaches which perceive administrative bookkeeping data as indications of ongoing processes. Previous analyses of process-produced data usually treated them in terms of the administrative purposes for which they were collected and rarely reinterpreted them as indicators of other events and processes. Basically however, nothing stands in the way of such an analysis of process-produced data. The prerequisite however, is further development of a new, social-scientific source criticism, which can adequately evaluate process-produced data in terms of the context in which they were originally collected.

<sup>51</sup> See: Bick, Wolfgang, Müller, Paul J. and Reinke, Herbert, QUANTUM Dokumentation 1978, Historische Sozialforschung, Historical Social Research, HSF, Vol. 5, Stuttgart 1978, p. XXIV, Bick Wolfgang, et al., Quantitative History in Transition, in: Social Science Information Vol. 16 (6), pp. 697–714.

## Appendix 1: Administrative Tasks

1	Einbürgerung	34	Eingliederungshilfe f. junge
2	Namensänderung		Zuwanderer
3	Staatsangehörigkeitsfeststellung	35	einmalige Beihilfe aus dem Härte-
4	Aufenthaltserlaubnis		fonds des Landes
5	Anmeldung d. Wohnsitzes	36	Sozialhilfe
6	Einzelhandelserlaubnis	37	Leistungen nach Unterhalts-
7	Versichertenrente		sicherungsgesetz
8	Reisegewerbekarte	38	Armenrecht
9	Reisepaß	39	Pflegehilfe als offene Sozial-
10	Fahrerlaubnis		hilfe
11	Kfz-Zulassung	40	Reisenkostenbeihilfe für DDR-
12	Bußgeldbescheid		Besucher
13	Leichenpaß	41	Beihilfe aus Anlaß von DDR-
14	Erlaubnis Gaststättengewerbe		Besuch
15	Erwerb d. dt. Staatsangehörigkeit	42	Hilfe für Sehschwache
16	Ausstellung e. Lohnsteuerkarte	43	Winterfeuerungs-/Weihnachtsbeihilfe
17	Anzeige über den Beginn eines	44	Zusatzreisen f. Besucher aus DDR
	Gewerbes	45	Ausweis f. Vertriebene u.
18	Personalausweis		Flüchtlinge
19	Unfallversicherung	46	Leistungen nach Häftlingshilfe-
20	Wehrerfassung		gesetz
21	Geburtsanzeige	47	Erholungsmaßnahme f. Kriegs-
22	Todesanzeige		opfer
23	Aufgebot	48	Minderung der Erwerbsfähigkeit
24	Bauantrag	49	Aufwendungsdarlehen
25	Kriegsopferfürsorge	50	öffentl. Mittel f. Bau von Miet-
	(erg. Hilfe z. Lebensunterhalt)		wohnungen
26	Tbc-Hilfe	51	öffentl. Mittel für Bau von Eigen-
27	Übernahme von Krankenhaus-		heim
	kosten	52	Wohngeld: Lastenzuschuß
28	Blindengeld	53	Wohngeld: Mietzuschuß
29	Erholungs-/Badekur	54	Wohnungsvermittlung
30	Kriegsopferfürsorge (einmalige	55	Ausbildungsförderung
	Leistung)	56	Kranken-/vorbeugende Gesund-
31	Erziehungsbeihilfe nach BVG		heitshilfe für Kinder
32	Wohnungsdarlehen für	57	Adoptionseinwilligung der Eltern
	Behinderte	58	freiwillige Erziehungshilfe
33	Übernahme von Rückführungs-	59	Erziehungshilfe — Heimunter-
	kosten f. Deutsche		bringung

60	Amtspflegschaft. -vormundschaft		Schwerbeschädigte
61	Übernahme des Kindergartenbeitrags	78	Krankenschein für Verfolgte
62	Mahlzeitendienst f. alte Bürger	79	Kindergeld
63	Adoption eines Pflegekindes	80	Arbeitslosenhilfe
64	Pflegegeldzahlungen	81	Arbeitslosengeld
65	Aufnahme eines Pflegekindes	82	Fortbildung/Umschulung — Unterhaltsgeld
66	Schulgesundheitskarte	83	Konkursausfallgeld
67	Untersuchungs-, Fürsorgebogen	84	Berufsausbildungsbeihilfe
	Tbc-Hilfe	85	Arbeitsberatung/ -vermittlung
68	Anmeldung AOK	86	Förderung d. Arbeitsaufnahme; Reisekosten, Umzug
69	Medizinische Leistungen/ Rehabilitation	87	Berufliche Rehabilitation
70	Schadensbericht	88	Förderung d. Arbeitsaufnahme; Überbrückungsbeihilfe
71	freiwilliger Beitritt AOK	89	Förd. d. Arbaufn.: Arbeitsausrüstung
72	Prüfung des Versicherungspflichtigen	90	Förd. d. Arbaufn.: Bewerbungskosten
73	Fragebogen d. Vollstreckungsbehörde AOK	91	Jobvermittlung
74	Krankengeld	92	Lohnsteuerjahresausgleich
75	Haushaltshilfe		
76	Kassenkur		
77	Bundesbehandlungsschein f.		

## Appendix 2: Personal Characteristics/Attributes

- I Personal data*
- 1 name
  - 2 address
  - 3 place of birth
  - 4 date of birth
  - 5 nationality
  - 6 sex
  - 7 telephone
  - 8 distinguishing marks
  - 9 other names, change of names
  - 10 marital status
  - 11 residential history
  - 12 dates of marriages
  - 13 dates of deaths

- II Family and household*  
14 relatives: in general  
15 obligation to support  
16 relatives within household/flat/dwelling  
17 child status

- III Residence*  
18 size of flat/dwelling, number of rooms  
19 equipment/state of flat/dwelling  
20 kind of utilization of flat  
21 landlord, hirer  
22 subsidized/non subsidized flat/dwelling

- IV Employment*  
23 practiced occupation (current or last)  
24 occupational status  
25 employer (current or last)  
26 employments: dates  
27 employments: other  
28 additional occupation  
29 occupational history  
30 end of employment  
31 periods of unemployment  
32 periods of being unfit for work

- V Education/qualifications*  
33 education  
34 professional training  
35 other qualifications

- VI Economic situation: income*  
36 income from self-employment  
37 income from employment  
38 income from subsistence payments  
39 annuity, pension  
40 income from subletting  
41 other income  
42 income (global)  
43 property/unearned income  
44 estimated constancy of income

- VII Economic situation: expenditures*  
45 business expenditures  
46 personal allowances



- 47 deductions for exceptional circumstances
- 48 flat/dwelling, real estate
- 49 debts, mortgages, loan, outstanding taxes
- 50 expenditures (global)

#### *VIII Health*

- 51 health data
- 52 medical history
- 53 links to hospital/doctor

#### *IX Memberships*

- 54 religion
- 55 voluntary associations

#### *X Military*

- 56 military

#### *XI Special population groups*

- 57 social problems: limited characteristics
- 58 social problems: economic aspects
- 59 social problems: social aspects
- 60 exiles, evacuees, refugees, „Lastenausgleich“, etc.
- 61 orphan/widow(er) of war, disabled/missed persons, „Spätheimkehrer“, etc.
- 62 physically disabled persons
- 63 tuberculosis
- 64 others: foreigners, political persecutees

#### *XII Links to the public order administration*

- 65 office for public order, register office
- 66 court of justice, bankruptcy, criminal procedure
- 67 divorce
- 68 guardianship
- 69 labour court

#### *XIII Links to the social welfare administration*

- 70 office for welfare/for the youth
- 71 labour office
- 72 social insurance for old-aged, survivors, etc.
- 73 social insurance for disabled, handicapped persons, etc.
- 74 juvenile detention centre, house for the aged, etc.

#### *XIV Links to other organizations*

- 75 building administration
- 76 revenue office

- 77 health insurance
- 78 banks/saving banks
- 79 global

*XV Other information*

- 80 data on motor vehicles
- 81 references, witnesses
- 82 wishes, intentions
- 83 arguments for private decisions
- 84 arguments for applying
- 85 opinions of other offices

### Appendix 3: Categories of Persons

- |                                      |   |
|--------------------------------------|---|
| 1 applicant                          | 15 children under 18 years              |
| <i>persons within the household</i>  | 16 children above 18 years              |
| 2 spouse                             | 17 stepchildren under 18 years          |
| 3 children under 18 years            | 18 stepchildren above 18 years          |
| 4 children above 18 years            | 19 parents                              |
| 5 stepchildren under 18 years        | 20 stepparents                          |
| 6 stepchildren above 18 years        | 21 grandparents                         |
| 7 parents                            | 22 grandchild                           |
| 8 stepparents                        | 23 brothers and sisters                 |
| 9 grandparents                       | 24 other relatives                      |
| 10 grandchild                        | 25 other persons                        |
| 11 brothers and sisters              | <i>special categories of persons</i>    |
| 12 other relatives                   | 26 children above 18 still in education |
| 13 non-relatives                     | 27 foster child                         |
| <i>persons outside the household</i> | 28 proxy                                |
| 14 spouse                            | 29 foster-parents, guardian             |

## Appendix 4: List of Municipal Authorities in the City of Nuremberg, 1970

- 1 Office of the first mayor (political head of the city government)
- 2 Office of the second mayor (deputy and administrative head of city government)

### *Divisions*

- 3 Division of Administration
- 4 Division of Finance
- 5 Division of Public Order
- 6 Division of Schools and Culture
- 7 Division of Social Welfare and Health
- 8 Building Division
- 9 Division of Public Utilities, Transportation and Industrial Development

### *Offices*

- 10 General administration office
- 11 Procurement office
- 12 Organization office
- 13 Office for central data processing
- 14 Personnel office
- 15 Disciplinary rules office
- 16 Office for urban research and statistics
- 17 Auditing office
- 18 Office for the budget
- 19 Treasurer's office
- 20 Tax office
- 21 Office for landed property
- 22 Office for legal affairs
- 23 City police
- 24 Office for public order
- 25 Registration office
- 26 Register office
- 27 Office for social security
- 28 Fire department
- 29 Office for civil defense
- 30 School board
- 31 Office of public culture
- 32 Art gallery
- 33 City archive
- 34 Central library
- 35 Theatre

- 36 Zoo
- 37 Public libraries
- 38 Education center
- 39 Museum
- 40 Schools
- 41 Welfare office
- 42 Dependency benefits office
- 43 Nursing office
- 44 Care for the elderly office
- 45 Office for the youth
- 46 Health office
- 47 Office for chemical analyses
- 48 Office for hospitals
- 49 Office for compensation (refugees, victims)
- 50 Office for sport
- 51 Board of works
- 52 Office for commissioning of public works
- 53 Office for urban planning
- 54 Surveying and mapping
- 55 Building laws
- 56 Office for surface building
- 57 Office for underground building
- 58 Office for public parks
- 59 Office for housing and settlement
- 60 Sanitation department (street cleaning, sewage disposal etc.)
- 61 Office for veterinary services
- 62 Office for market regulations
- 63 Office for public baths
- 64 Office for funerals

## Priorities for Record Linkage: A Theoretical and Practical Checklist

Historical record linkage, as S. Langholm has recently noted, is „a whole little science if its own“. By this term we generally mean the bringing together of historical records relating to the same historical unit of interest — usually a person, but often as well, a family, a household or perhaps a process, event or object. And usually when we do this we employ routinely generated records such as parish registers.

Although this little science (or sub-science) is a new one, it is logically connected with a very much older historical practice, that of saying more than one thing about historical individuals (units) by means of reference to more than one document or record relating to that historical individual or unit. Although the logical part of this little science is shared by all historical activity (the part connected with the reidentification and further characterization of historical individuals), the technical parts are of greatest interest to the field of micro-history and its blood-brother, micro-demography. In both of these fields we are primarily concerned with building up collective biographies of individual people, or of individual families, or of individual households prior to our analysis of the resulting data. As a rule, the main sources employed for such purposes have been parish registers (France, England, Sweden, Denmark) or, for the 19th century especially, census rolls (U.S.A., Canada, Britain) and their near neighbors taxation lists or assessment rolls.

A more recent practice has come to be the use of either of these more or less total record sources as a backbone for the research in question, with the addition of as many other sources as well. In principle, there is no limit to how many record sources one might use.

Since interest in the micro-historical and micro-demographic fields is both high and increasing, and since many more individuals and groups are considering using record linkage practices in their work, I want to discuss problems of priority. That is, I want to talk about those general problems which anybody engaging in micro-historical or demographic work may face in a record-linkage context. The specific problems, while they have not been well reported, have at least been reported. And this work is readily available to the would-be record linker. In this regard I mention a book edited by E. A. Wrigley called *Identifying People in the Past*, an article by Theodore Hershberg in the *Historical Methods Newsletter* in 1976 and an earlier article by myself in the first issue of the *Journal of Interdisciplinary History*<sup>1</sup>. I shall

<sup>1</sup> Wrigley, E. A. (Ed.), *Identifying People in the Past*, London 1973; Hershberg, T., et al., Re-

assume that if you are really interested in this area you will read these few things. What the literature really lacks is a discussion of the variety of factors which face a researcher or research team who wish to know whether or not they should automate, partly automate or do the entire job by hand. And, furthermore, whether or not they should perform a sample linkage first, whether or not they should establish an iron-hard set of linkage rules in advance, and what general problems they should consider and possibly tackle with some zest.

## 1. Factors Affecting Record Linkage Priorities

The would-be micro-historian or demographer will find his record-linkage planning constrained by three obvious, but usually not well thought out, features of his task. First, by the resources that are available to actually bring about the necessary record-linkage. And, secondly, by the problems posed by the files with which he must work. In Figure 1 I have listed these two headings and their main sub-headings.

*Figure 1: Factors Affecting Record -Linkage Priorities*

---

Available resources	<div data-bbox="441 776 829 862" style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;"> Staff available (pay/no pay)  Data processing (free/charge)  Money (quantity/years avail./strings) </div>
Problems posed by files to be linked	<div data-bbox="441 873 974 1130" style="border-left: 1px solid black; border-right: 1px solid black; padding: 5px;"> Desired organization of final file  Number of files to be linked  Size of files  Condition of files  Kind of linkage (total/sample)  Available identifying items (common/overlap/independent)  Biases  Legal or moral constraints </div>

---

Each of the above factors and their sub-factors are interrelated. And each potentially affects the choice of record-linkage priorities. The basic outcomes of any such consideration of record-linkage priorities are decisions as regards a) file organization, b) file preparation, c) the linkage steps themselves. As regards the first of these,

cord Linkage, in: HMN, 9 (1976), pp. 137–163; Winchester, Ian, The Linkage of Historical Records by Man and Computer: Techniques and Problems, in: The Journal of Interdisciplinary History, 1 (1970), pp. 108–124.

we have two basic decisions to make: How to organize the files prior to the linkage; and how to organize the master file after the linkage step or steps. File preparation refers to how we wish to prepare the files in advance of the linkage step(s) — for example, by means of a number of data transformations to enable detailed comparisons of records to be made more easily. The major decisions to be made with respect to the record linkage steps are, first, whether to do the entire process by hand, or to enlist the aid of some data processing device(s) for part or all of the actual linkage, and second, what exact linkage rules and steps to employ.

Before going on to say a few things about the items in Figure 1 and their relation to the three basic kinds of decisions, I shall give, in Figure 2, a chart summarizing the latter.

*Figure 2: Possible Outcomes of Record-Linkage Priority Considerations*

---

Decision as regards

- |                             |   |                 |                             |      |
|-----------------------------|---|-----------------|-----------------------------|------|
| a) file organization e. g.  | sort n-files alphabetically<br>organize output files by individual<br>family<br>household<br>parish   |                 |                             |      |
| b) file preparation e. g.   | delete no information<br>delete all nominal information<br>recode all surnames<br>recode all given names<br>classify all occupations  |                 |                             |      |
| c) linkage steps            | 1. determine exact linkage rules<br>2. choose among <table border="0" style="margin-left: 20px;"> <tr> <td>fully automated</td> </tr> <tr> <td>partially automated linkage</td> </tr> <tr> <td>hand</td> </tr> </table> | fully automated | partially automated linkage | hand |
| fully automated             |   |                 |                             |      |
| partially automated linkage |   |                 |                             |      |
| hand                        |   |                 |                             |      |
- 

In terms of these various factors and outcomes which affect record-linkage choices, there are two orders of interest: an order of importance, and an order of temporal priority. For example, it is a mistake to determine exact linkage rules prior to the examination of the files under consideration. And it might be a mistake to decide prior to such a consideration that one will opt for a fully hand-linked operation. So, simply because the linkage steps might be a paramount, or a high priority consideration, it does not follow that they should be determined first.

Probably the most important consideration of all is what the historian/demographer/geneticist/etc. intends to do with the linked files. Usually the answer is clear. The researcher wants to create cross-tabulations of certain kinds which form the basis of his descriptions or explanations, and which may suggest other forms of analysis or other problems. If this is so, then the first matter to which consideration should be given is the matter of the desired logical organization of the final or out-

put file. This could be in the form of a more orderly collating of separate cards full of information or of their card-image equivalents. But if the studies to be undertaken involve detailed examination of, say, married siblings or of multi-generational family groupings, then detailed consideration will have to be given at an early stage to the graphical or list representation of the output file.

Perhaps the second most important consideration, next to that of knowing roughly what you want to do with the linked files, is that of getting acquainted with the files in a detailed way. It is really crucial at an early stage to get to know the quirks, the strengths and weaknesses and biases, of one's files. There are exceptions to this rule. But it is really very important for the would-be record linker to plan a honeymoon with each of his files in turn prior to linking them, rather than to couple them together into a harem and then imagine that detailed familiarity will come more easily later. None of the record linkage decisions listed in Figure 2 are possible unless a researcher is acquainted in a technical way with his files prior to any linkage on a big scale. The sort of technical way I have in mind is detailed in the articles and book I mentioned earlier. But in summary they are these:

What identifying or potentially identifying items are common to or overlapping on the files to be linked?

With what relative frequencies do the items occur? (e. g. Are there more Browns than Schmidts and more Russells than Johannssons?)

What varieties of names are there with similar spellings, translations (White-Blanc-Bianco), or transformations because of dialect, patois or identifying necessity?

With what frequencies are identifying items likely to be discrepant on linked pairs of records?

Exactly how large are the files? What is the distribution of the various identifying item frequencies in each item kind? What is the range of such frequencies?

If one has this kind of information about one's files at an early stage, thinking about file preparation, organization and the detailed linkage steps is made much easier. This ease is partly because of the mere fact of detailed familiarity. But it is also because each of these other matters requires decisions based on the kinds of knowledge referred to above. Most of us who have set up rather large data bases involving a linkage component have stumbled into the matter without such prior detailed considerations of our files. We did not know that it was needed, nor could we imagine before the fact why such detailed consideration would be needed. If one has such knowledge at an early stage, then one can use that knowledge for the following purpose:

To estimate the time required to complete a particular portion of the linkage job.

To estimate the cost of completing a linkage job by hand or by machine or partly by both means.

To estimate biases in one's data or in potential sampling procedures or in potential linkage procedures.

To determine how to prepare the files as regards standardization, transformation of identifying items such as names into a standard format for sorting or detailed comparison purposes.

To estimate the relative weights to be placed on the agreement or disagreement of various identifying items during a linkage step.



And therefore:

To determine, or help determine, the exact linkage rules needed.

And finally, because of all these things:

To help the researcher choose among a fully automated, partially automated or completely un-automated linkage procedure.

The third most important consideration is that of knowing as exactly as one can (or being as scrupulously honest as one can) about the matter of available resources for linkage — including file preparation, detailed planning, programming and supervising in terms of money, people or free computer time or free programming help. This assessment is an enormously important consideration. But it ranks third in my suggested list because I am an optimist who in his heart believes that where there is a will a way can be found. It also ranks third because, if you do not have any idea about the problems you want to tackle first and if you are not acquainted in detail with the files you want to use in the manner I have suggested, then you cannot begin to know if the resources you have available are enough to enable you to do the job. There is room for a little circularity here, of course. For if you do not know that you have some resources, then you cannot plan even to have the detailed familiarity with your files that I suggest you have. Here, however, I think that the historian's traditional resource (himself and perhaps a graduate student or two) will probably suffice in a pinch.

I cannot really hope to give much useful advice here on the matter of comparing needs to the resources available. So I will offer, instead of something theoretically and practically satisfactory, a comparison of a number of successful projects and of the resources which they have available relative to the tasks they are pursuing.

I have chosen five projects with their record-linkage tasks. The five I have chosen cover a fair range of the spectrum of tasks and problems, as well as a fair range of file sizes and problems. In the order of their reporting of their record linkage techniques and problems they are: The Hamilton Project, The Cambridge Group, the Philadelphia Social History Project, the Umea Demographic Database, and, in order to illustrate small files, Stewart Hardy's Model School Study.

The Hamilton Project has reported some of its methods and problems<sup>2</sup>. The Cambridge Group have published a long discussion in the book edited by Wrigley mentioned earlier<sup>3</sup>. The Philadelphia Social History Project has recently published an article which builds mainly on Winchester's 1970 article, but contains a very interesting discussion of bias using census to census linkages<sup>4</sup>. The Umea Demographic Database has reported on its files in detail. There is no discussion of record-linkage as such in these reports, because, I believe, there is a belief that there are no

<sup>2</sup> Winchester, *op. cit.*

<sup>3</sup> Wrigley, E. A., and Schofield, R. S., *Nominal Record-Linkage by Computer and the Logic of Family Reconstitution*, in: Wrigley, People, pp. 64–101.

<sup>4</sup> Hershberg et al., *Record-Linkage*.

problems<sup>5</sup>. If true, this is a remarkable fact. Finally, Hardy is a doctoral candidate who has nearly completed his work on Model Schools in Ontario in the last century. His work exemplifies small, multiple file linkages<sup>6</sup>.

What we want specially to dwell on is the match between the resources available and the magnitude of the task faced by each researcher or team.

Both the Cambridge Group and the Demographic Database work with parish records. The Hamilton Project and the Philadelphia Social History Project work with 19th century census rolls and other parallel sources. If I were to rank them in order of size in terms of the sheer data which the projects handle the order would be:

Demographic Database	Gigantic	Approx. $10^7$ records
Philadelphia Project	Very Large	$10^6$
Cambridge Group	Large	$5 \times 10^6$
Hamilton Project	Medium	$10^5$
Model School Project	Small	$3 \times 10^3$

where by a „record‘ I mean an „80-column card image“.

In terms of the difficulties which the data pose, however, I would give quite a different ranking, rather more like:

Cambridge Group	Extremely difficult	only names, many variations, bad handwriting
Hamilton Project	Very difficult	25 % of items with discrepancy on matched pairs, handwriting difficult
Philadelphia Project	Very difficult	similar to above
Model School Project	Fairly difficult	only names and ages (sometimes)
Demographic Database	Fairly easy	often whole families, much linkage done by priest at time

In terms of the resources available, since these are projects in progress or completed, there is a direct correlation between the resources available and the size of the project files. My best guess as to the ranking, personnel, money available and computing power is as follows:

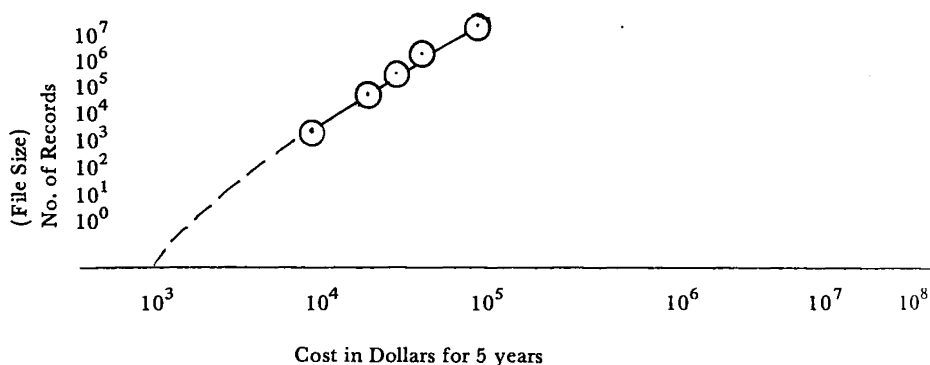
Database	Personnel	5-year funds	Computer power
Umea DD	Clerical 40 Technical 1 Supervis. 2 Historian 1	$\$1.5 \times 10^6$	Tape to disc IBM 360 No extra cost

<sup>5</sup> Johansson, Egil, and Sundin, Jan, The Demographic Database, I and II, mimeo. Umea 1977.

<sup>6</sup> Hardy, Stewart, Linking Educational Records to a Manuscript Census, mimeo. Ontario 1977.

Philadel.	Clerical 4-12 (part-time) Technical 2 Sup./Hist. 1	\$ $5 \times 10^5$	Card input Interactive setup IBM 360 Extra cost
Cambridge	Clerical 2 Technical 1 1/2 Historical 3	\$ $2.5 \times 10^5$	Paper-tape input Nottingham IBM 360? Extra cost
Hamilton	Clerical 1-2 Technical 1/2 Historian	\$ $1.25 \times 10^5$	Card input IBM 7094/360 No extra cost
Model Sch.	Historian 1	\$ $2.0 \times 10^4$	IBM 360 & Minicomputer No extra cost Hand calculator

If we were to graphically display the relationship between the funds and the file size it would look something like this:



Of course, the costs I have graphically displayed in comparison with the file sizes are the total costs of running a viable historical research project. They are not just the costs of recording the data in a usable form (except for the Umea project). It would be possible to project costs going the other way — namely from unit costs estimates and size of job to costs for a project of a particular size for a five year period. But I think that these rough approximations to reality suggest quite clearly how expensive a large database using linkage techniques is.

It is probably unfair to try to do a ranking of productivity in terms of scholarly product of the various projects listed. But it would be reasonable to expect that projects with a higher number of historical-academic staff would be more productive in a given five year period. In this regard, I will invent as a publishing unit a „Russell“ which is the rate at which Bertrand Russell published in a year . . . roughly one book and ten articles, for a total of perhaps 500 pages of print. At this rate, the Cambridge project has been publishing at roughly a Russell/year or more, for a total

of possibly six or seven Russells in a five year period. The Hamilton Project is pretty close behind in the period 1970–1975 with five annual reports of some 300 pages each, a number of independently printed articles by Katz and Winchester and a book by Katz for possibly a 4.5 to 5 Russell total. The Philadelphia Project in the period 1972–1977 is loping along at about half a Russell a year, but is picking up the pace rapidly in the last year or so for a five year total of some 3 Russells. However, recently a flood of scholarly workers is on the scene. The Umea project has put almost all its effort into the data transcription, file preparation and linkage process in the past five years. Egil Johansson's work on literacy has been the main research using the database to date. And while this seems to me to be the most significant, perhaps, of all the work done, it does not amount to more than a Russell or two all in all. Hardy's work has so far issued in a couple of papers<sup>7</sup>.

The point of this aside is absence of necessary correlation between effort in studies involving record linkage and either quantity or quality of results. Consequently, while it is certainly necessary to see that the total resources one has available are sufficient for the task at hand, I think it is still of higher priority to know what one wants to accomplish with the research in question and to be acquainted with one's sources in intimate detail. Then the resources can be found, or often can be. Or else one can tailor one's research to the resources one has. But to do this requires that the resources and the technology be made a third-rank priority.

So in summary, I would list the priorities for the record-linker to be as follows in rank order:

1. Knowledge of what is to be accomplished, of the kinds of problems to be tackled;
2. Intimate knowledge of one's source files;
3. An honest appraisal of one's five year resources.

I am strongly recommending, therefore, that the concepts and problems come first and the technology a distant third.

## 2. To Automate or Not to Automate, That is the Question

Having given my global priority recommendations I will now try to say some words about the vexing problem of when to commit oneself to help by a machine and when not do so. Three rules of thumb suggest themselves to help one decide whether to link by hand or to attempt to automate the linkage. These rules are as follows:

1. If the files are very messy and the identifying items few, then one should link by hand.
2. If the files are very large, and the identifying items more than adequate, then one should link by computer or partially so.
3. If the files are small, then one should always link by hand.

<sup>7</sup> Hardy, Stewart, Educational Records.

As general rules of thumb (as opposed to general rules) these three are pretty good. But they are certainly inadequate to cover all cases. I shall attempt to discuss those cases in which they are inadequate, because it is here that mistakes can be costly of time and effort.

In the case of the first rule, a great deal hinges on the notions 'very messy' and 'few identifying items' which are not, as they stand, very scientific. Even if there are 20–25 % errors in the identifying items which one has, there exist perfectly good methods for reducing the potential effect of these so that one can by automated means achieve roughly the same result as by an intelligent filing clerk. I shall not go into details here since the matter is pretty well covered in the literature<sup>8</sup>. The basic idea here is simple. If, for example, names relating to the same individual have a certain tendency on the average to be differently spelled, one can counteract that tendency by systematically transforming the names on the files to be matched to a kind of common spelling. (For example, Smith, Smyth and Smythe might be transformed to SMTH.) Similarly, if other potentially identifying items (such as age, for example) are often in disagreement when two records relate to the same person, and if we can estimate the frequency and the extent of disagreement, then we can devise data transformation schemes to compensate. This method, of course, presupposes a familiarity with one's files of the sort which I mentioned before — very likely in the form of a small hand-linkage study of them. Suppose, for example, that you find that ages can often be discrepant by as much as five years on pairs of records that relate to the same person. And furthermore, suppose that it is not uncommon that the figures in the ages are transposed in the inscription: '25' and '52'. Then we need, at the time we come to compare identifying items on a pair of records we are considering for linkage, two transformation rules. One is that of considering the age on one record and comparing it with that on the other in such a way that we consider it evidence for the linkage if the ages are within five years of one another. The other is that of comparing the ages and their digit reversals. If one is the reversal of the other, then we consider this evidence in favour of a link. If the frequencies of either of these occurrences are known or can be estimated from a sample linkage, then we can weigh the evidence<sup>9</sup>.

I would, myself, follow 1. and do the linkage by hand if I thought that the files were so messy that no sample of them would enable one to predict what was coming in the next batch. And I would similarly do the linkage by hand if my computing and clerical resources were minimal. Otherwise, I think one should do as much of the work by mechanical or electromechanical means as one can.

<sup>8</sup> See, e. g., Winchester, *Historical Records and Hershberg, Record-Linkage*.

<sup>9</sup> See Winchester, *op. cit.*

## 2.2. Large Files

It seems obvious that if the size of files is very large and the identifying items more than adequate, then one should link by computer. I still think this a pretty good rule of thumb. But I no longer think that it is a general rule that one abandons at one's peril. There are two factors which can throw this obvious rule into doubt. The first is lack of computing and programming power. And the second is excess of clerical power. If you do not have a computer available, preferably free, with adequate storage facilities, large enough core capacity, then you would probably do better to take a decade and do the job by hand, writing articles as you go. This method is certainly better than trying to manage on an inadequate budget with an inadequate machine. Another ground for doubting the wisdom of this advice is if you have foreknowledge that in a year or two the machine you have (and the programmers or systems people) is going to be obsolete, or transformed into a minicomputer, or sold. It seems to me better to take five years out of your life and have you and your graduate students do the job by hand than to constantly have to reprogram or become acquainted with new machines, people and their quirks. This situation is a general problem recognized in the computer industry under the general heading „the interface problem“. And it is a genuine problem and a very great bore for historians — though, of course, a lot of fun for computing people. Of course, if you have the money, the computer and adequate programming backup, then by all means automate or semi-automate.

If you have an excess of clerical staff, then you may do well to let others do the automating. This course of action requires, for very large files, that a number of other conditions be met. The only place where I know that this has successfully been done is at the Demographic Database in Umea, Sweden. The files being used in this case are beautifully kept parish records from the 19th century which are systematically linked to other registers maintained by the clergy from the 1740's following a Royal Decree of 1748. Essentially a special register of the capacity of each household to read and write — by means of graded marks — is linked to vital statistics registers. Each time an individual appears in one of the five basic sources an excerpter fills in a card. At a later stage in the operation all of the cards relating to a single individual are sorted together by hand before being keyed to a disc storage device for later data processing by computer. Since the system is based on individuals, a so-called „guide-card“ is also produced by hand which gives, for every individual, his own identify, his father's, his mother's and the identify of the husband and wife. Links are thus made from son to father and between husband and wife, thus facilitating simple family reconstructions at a later stage.

Since there are forty clerical staff involved in this process, and since there is little difficulty in making the links for anywhere except Stockholm or between Stockholm and another parish, the linkage work can be done very quickly and at the same time as the original excerpts are made. Thus even though there are excellent computer facilities available in Umea, I see no reason why the linkage portion of this mammoth operation should be automated. Though, perhaps, when Stockholm is tackled some time in the next century one might want a little machine assistance!

### 2.3. *Small Files*

This commonsense rule of thumb runs „If the files are small, then link by hand“. And for two small files (say two voters lists in Cambridge in successive election years in the mid 1800's) this is certainly a good rule. But if more than two files are under consideration, the number of detailed comparisons between record pairs which one may have to face (especially in parish register work with only nominal identifying items) can become astronomical. The difficulty which one faces by hand-linking files which are essentially unordered is that one has, in effect, potentially to compare every record to every record in each of the other files. Even with only three distinct files with five records in each file the number of possible linkages is large, with the number of distinct persons possible ranging from five to fifteen. If (in the worse possible case) all of the names were similar to one another and one had a limited amount of information to enable the resolution of ambiguities, one would have to consider 12,962,661 possible linkage arrangements<sup>10</sup>. In the best of all possible worlds, one would still have to consider 125 possible links. And as the number of files increases, the number of links possible goes up exponentially.

Clearly, even in the three file case, something must be done — to bring the number of comparisons down to feasible proportions. The most obvious and most used strategy is to reorder the data in each file alphabetically or numerically and to limit the detailed comparisons between records (among records) to those within a limited range or „pocket“ within the sorting key. For example, with linkage mainly dependent upon surnames and forenames one might sort all files by surname and only consider for detailed consideration those surnames in each file which have the same surname. For very messy files, especially small ones, this can be a disastrous manoeuvre since one might lose twenty percent of one's actual links. Thus, again, one might be driven to some sort of coding of surnames which will bring together all of those surnames which are sufficiently similar to warrant a detailed comparison. If this sort of data transformation is too strict, it misses most of the links. But if it is too loose, one is back to comparing nearly all records with all records. Here there is room for art! Most of this sort of thing can be managed with only a card sorter and a keypunch machine for small files. But for many of these, a computer is a very useful aid. So while the rule of thumb is pretty good, it has its notable exceptions.

One final note under this general heading. If the files are very messy, the identifying items few and their quantities very large, then one would be wise to abandon the task. But since I am describing the situation facing those using parish registers in England and in many other countries, or censuses prior to 1840 pretty well everywhere, perhaps my best advice would be to befriend an oil sheik who has a passion for microhistory and demography.

<sup>10</sup> Skolnick, Mark, *The Resolution of Ambiguities in Record-Linkage*, in: Wrigley, People, pp. 102–127.

### 3. Definite Linkage Rules for All Files

I would now like to offer a rule of thumb which I think should be given the dignity of a general rule in present day historical practice. *That is, all linkages, whether small or large, with messy or clean files, with computer or by hand or partially automated, should follow definite rules which are reported as a standard part of the project.* This is a rather long and somewhat priggish rule of thumb. It is a recommendation for which I have propagandized in support of both with my students and with anybody else who sought my advice on their record linkage problems.

My prime reason for this bit of stubbornness is this. It is only if one knows in advance the linkage rules which one is applying, that one can consistently link records such that one's linkage can be reproduced by another scholar. Now, just as in chemistry one should rightfully specify one's methods and processes so that they can be reproduced by someone else, so should one do the same thing in history. Such a simple step as this would go a long way towards increasing the respect for microdemographic or microhistorical research among other scholars. Certainly the beauty of one's descriptive prose need not be affected by such a step. And the plausibility of one's argument might be increased considerably.

Furthermore, although such a simple reporting of one's record linkage rules would enable another scholar or scholars to reproduce and check one's work, the result would likely be quite the reverse. Since one could check another's work at any time, one need not. Trust begins in shared methods. Some of the most strident controversies of the last number of years involving 19th century American cities and their social structure have arisen partly because of an inadequate reporting (as well as thinking through) of methods.

There are also reasons for having such definite linkage rules formulated and followed, even for the benefit of the scholar who is himself using them. The most obvious reason is the ease and comfort which systematization brings in its train — basically a gain in confidence and clearheadedness. But the reason of more scholarly, rather than psychological, import is that it will enable the estimation of systematic biases.

Of course historical sources of the routinely generated kind are always biased. Ecclesiastical parish registers are denominational. Thus non-Anglicans in England, non-Catholics in France will either be missing or underrepresented in parish registers. The only exception I am aware of is that of an established church which tolerates, for statistical purposes, no exceptions — such as the Lutheran Church in Sweden between 1640 or so and 1900. Tax registers tend to ignore the poor or to cover them less fully. Differences of sex, marital status, age and the length of residence, or quality of health, may all lead to differential coverage in the kinds of records which we may wish to link.

But record linkage processes, as such, can compound such biases or create new ones. The crucial issue for our discussion here is what kinds of biases record linkage processes do or may add to the initial circumstances. Now as a rule, and excluding



the Swedish work alluded to above, any record linkage process can be seen as a sampling process in practice.

This process can be seen by the following illustration. Suppose we wish to compare the incomes with the family sizes of the various scholars who engage in record linkage practices. (Perhaps we have a theory about the effect of such practices both on income and on fertility!) Our first file looks like this:

<i>File A</i>	<i>Incomes of Record Linkers</i>	
Surname	Given Names	Income (Annual)
Henry	Louis	250,000 fr
Wrigley	E. Antony	5,000
Schofield	Roger	4,500
Newcombe	Howard B.	30,000
Felligi	Ivan	47,000
Winchester	Ian	25,000
Hershberg	Theodore H.	100,000

and the second like this:

<i>File B</i>	<i>Family Size of Record Linkers</i>	
Henry	Louis	27
Sunter	Ian	2
Kennedy	James M.	7
Skofeld	Roger	2
de Wyncestre	Iain	1
Johansson	Egil	3

I think that, using all our wiles, we might come up with the following result for these two files, namely:

*Linked File: Incomes and Family Size of Record Linkers*

Henry (Henri)	Louis	250,000	27 children etc.
Schofield (Skofeld)	Roger	4,500	2
Winchester (de Wyncestre)	Ian (Iain)	25,000	1

Our problem as historians using record-linkage techniques is how to use this sample, generated by the vagaries of the record-linkage process, as representative of the entire lot of record linkers. If the sample is random, we can do quite a lot. But if there are systematic biases in our files of which we are not aware, then we stand the risk of producing a systematically biased, non-random sample — one which we have no clear way of handling.

Furthermore, if we are dealing with very large populations, then we may wish to sample the files to be linked in advance. The net result of the record linkage process could be the sampling of a sample. And depending upon the initial sample chosen it could be a biased sample of a biased sample. I do not know of any discussions in

the statistical literature which exactly parallel this process. Felligi and Sunter<sup>11</sup> suggest ways of handling intercorrelations among identifying items which may destroy the randomness of the linkage process, central to their mathematical model of the process. But what we are talking about are biases due to the historical, rather than the statistical, qualities of the data.

The problems can best be illustrated by two of the tasks which have been central to North American Research in the last decade. The first is the study of „persistence“ or its inverse „population turnover“. And the second is the study of occupational, geographical and economic mobility.

In persistence studies, the proportion of a given population persisting over time has been tied to a variety of significant historical features. Persistence can function both as something to be explained and as something which, if given, explains other features in the data. Thus, if rates of persistence are known for various communities, we can inquire how these are affected by differences among such communities, such as size, location, history, age, economics, rates of growth, population composition, access to cheap transportation and the like. As an independent variable, degree of population turnover has been used to explain all of the following<sup>12</sup>: lack of class-consciousness, limited working-class militancy, the slow growth of labor organizations, community instability, a variety of social pathologies, the continuing control by a small social elite and a stable social structure. There are two problems connected with record-linkage processes here. The first is that for any such persistence study we need to be reasonably sure that our rates of persistence are neither gross overestimates nor gross underestimates. And estimated rates are a function of the nature and manner of the recording of the data, the quality of the data for purposes of linkage and the exact record-linkage algorithm used.

The second problem is whether or not the qualities of those who persist differ significantly from those who do not. If they do not, then it is the mere fact of persistence in the community which is under study. But if they do differ, then it is reasonable to inquire into the special causal circumstances which may obtain. Again, in judging such research, both one's own and that of others, the quality of the reporting of the record-linkage processes is crucial.

In social mobility studies another feature connected with the record-linkage algorithm used is important. Whereas in persistence studies we want to use all the variables we have available which can potentially function as identifying items, in mobility studies such use can be a source of systematic bias. Suppose, for example, one wishes to follow occupational change through time. If occupation is also used as an identifying item — and it is a very good one — then we might potentially reidentify a higher proportion of people whose occupation did not change or which changed marginally during the time period under consideration. The result would be a biased sample produced by our process. Since such a bias is hard to avoid in

<sup>11</sup> Felligi, I. P., and Sunter, A. B., A Theory for Record-Linkage, in: *Journal of the American Statistical Association*, 64 (1969), pp. 1183–1210.

<sup>12</sup> Hershberg, Record-Linkage.

hand linkage studies, this is another reason why standard linkage practices and reporting of these would be a great boon to workers in the field.

One final example of potential bias produced by linkage processes: In order to make linkage easier a number of researchers have used a strategy of systematically deleting commonplace names from their files. This practice certainly does make re-identification much easier and considerably reduces the files under consideration. But it also means that if it is the common man (Schmidt, Smith, Jones, Andersson and Lefebvre) who is the object of study, then the most common of all is systematically left out. There has been some discussion of Thernstrom's use of this method<sup>13</sup>. Standardized reporting of linkage rules and methods would have at least facilitated or, more likely, made such discussions unnecessary.

It is, I think, clear that definite linkage rules for all linkages — hand, machine-assisted and computer-automated — are a boon to our art. And I think that it is also clear that we require some standard means of reporting on our files and our rules which we all can understand.

#### 4. Standard Reporting for Everyman

I shall not argue further here for a standard reporting of one's files and linkage rules for all studies involving record linkages in a historical context. What I want to do is simply suggest a minimum list of things which should be reported. To do this I shall refer back to Figures 1 and 2. I think that a footnote or an appendix to published work involving record linkage should mention at least the following:

##### A. Original Files:

1. The number and type of files linked.
2. The size of each file.
3. The organization of each file in its original form.
4. The number and kind of available identifying items.

If there are many files, then whether the identifying items are common to all or only by file pairs.

5. The condition of the identifying and descriptive items on each file. Emphasis should be placed on such things as surname variations and the likelihood of discrepancies in identifying items in truly linked record pairs.
6. Systematic biases in original files.

<sup>13</sup> Alcorn, R. S., and Knights, Peter R., *Most Common Bostonians: A Critique of Stephen Thernstrom's The Other Bostonians, 1880–1970*, in: *HMN*, 8 (1975), pp. 98–114; and Thernstrom, Stephen, *Rejoinder to Alcorn and Knights*, in: *op. cit.*, p. 117.

## B. Files as organized for linkage:

1. The preliminary preparation of each file based on 5 above. Whether or not re-coding of, say, surnames or occupations has been done. Whether a derived sorting key was added by hand or by a machine step. Whether surnames were given a standardized spelling prior to linkage steps.
2. The file organization prior to linkage. Whether this is the same as on the raw files (3 above). Whether files have been sorted according to some sorting key or other (e. g., surname, sound, by occupation code). Whether each file is organized by individual, family, household, parish and the like.
3. Whether the physical order of the files is the same as the logical order.

## C. Linkage Steps:

1. Whether the linkage is a hand operation, a partially automated one or a fully automated one.
2. The results of a preliminary hand linkage, if undertaken.
3. The exact linkage rules followed.
4. The final file organization for the linked file.
5. The biases which the particular linkage steps may involve.

If such a reporting became a standard procedure among micro-historians and micro-demographers, both the quality and the comparability of what we do would be improved significantly. I think we would also be more convincing.

It is time to summarize what I have said, both to suggest a number of unsolved and untackled problems and to mention some future possibilities which recent technology tends to make possible. I have tried, by way of filling a gap in the literature, to discuss a number of factors affecting record linkage priorities and decisions in the context of some work involving the techniques in the last decade. I have suggested that there are three overarching priorities for any would-be record-linker: namely, that he know exactly what his problems are and why he needs such a linkage; that he have an intimate knowledge of his source files before pushing ahead to plan the linkage steps; and that he be as honest and knowledgeable about his exact requirements and resources as he can be. If these three priorities are seen to, the researcher is pretty certain to prevail. I have tried to give some examples in detail as to what the range of possibilities might be in each case.

As regards whether or not to automate, I have suggested three rules of thumb, each of which has important exceptions. These were, first, that if the files are very messy and the identifying items few, then one should probably link by hand. But if the files are also very large and one has access to clerical and computer assistance, then techniques exist which can overcome the technical difficulties. If the files are very large, the second rule of thumb is that one should think of machine assistance from the beginning. But, if one has a large clerical operation available and the linkage step is a natural one, then one can avoid using a machine — especially if the files are clean and there is hardly ever doubt about a linkage. The third rule of thumb is that if the files are small, then one should always link by hand; but if one has many such

files, the potential number of searches one would have to do increases hypergeometrically. Thus in some cases machine assistance would be helpful.

As regards the matter of definite linkage rules, even for hand linkages I have argued that one should always have a precise set of linkage rules whether or not one is using a machine for any part of the operation. We are sloppy and easily deluded animals — even in our spiritual parts. Clio is a quiet and careful muse. Reporting one's record linkage procedures in a standard fashion would be a boon to our art. I have given a brief checklist of matters which such a reporting should, at a minimum, include.

For those interested in „the whole little science“ as a problem in itself, I have mentioned in passing a number of problems which we would do well to dwell upon in a technical fashion. We need a good discussion of biases and of sampling, which are creatures of the record linkage processes themselves. We need a much fuller discussion of the record linkage processes in terms of graph theory, especially since recent developments in computer science are tending to see graphical descriptions as important tools for describing many processes and files. We need somebody to draw all of what we know together into a handbook for historians and demographers, since what we have is scattered and incomplete. We need a series of standard programs for those using parish records and for those using other file types which we can use as easily as we use SPSS and Data-Text. We particularly need some interactive programs so that we can automate that which we can and can stop to look at specially difficult problems as the process proceeds.

Are there any interesting technical developments which might aid us in the future? Well, there is at least some hope that the central processors of the future will have much more storage to play with and will function at much higher speeds. But the most interesting developments, to my mind, will be in the training of young historians who will take all of what we are presently puzzling over as a natural and commonplace part of their education.

## Problems in Handling Process-Produced Data

### I. Introduction

I am assuming that the other contributors to this section on „Preservation, Storage and Assess“ will deal more specifically with the technical aspects of „Storage“ and „Preservation“ of large-scale data sets and, possibly, with the computing aspects of the problems of „Access“. I wish to address myself more specifically to a variety of problems in the area of „Access“ from the initial stage of acquiring information about Process-Produced Data, through the requirements for adequate documentation, to the final phases of actually analyzing Process-Produced Data sets.

Readers will note that most of the examples and illustrations given throughout this paper refer to demographic data, mainly from Canada. Although these illustrations reflect my personal experiences, biases and ignorance, I am sure that you can find similar examples from your own country and your own academic discipline. I am strongly convinced that, generally, the problems I will discuss are not unique nor are they particularly more serious in one country than in another. At best, one can state that some countries appear to have made somewhat greater progress towards solving the following sequence of problems than have other countries.

I should also note that there are a variety of definitions of „process-produced data“ used throughout the social scientific literature. A fairly cursory search through various journals yielded *five* definitions, which are not identical with each other. I will not enter into a detailed discussion of the respective merits of these definitions at this point; it is, however, obvious that sooner or later we will have to come to grips with this proliferation of terms.

In the following „tale of woes“ I construct the hypothetical sequence of problems which the researcher interested in process-produced data is likely to encounter during his search for and acquisition of administrative records as a data source.

\* I wish to point out that an earlier version of this paper has been presented at the First Open Conference of IASSIST (Canada) in Toronto, May 11–12, 1977. The earlier version was specifically aimed at IASSIST participants and thus dealt in greater detail with the kinds of actions required to overcome some of the problems which one encounters in handling process-produced data. Interested persons are invited to request copies of the earlier paper.

## II. Problems of Information and Acquisition

Under this heading, we can classify a large collection of problems which fall into several sub-categories:

### *A. Information about Existence*

To a large degree, information about the existence of particular process-produced data sets is rather hard to obtain. In fact, it is sometimes difficult to find out where such information may be obtained. Admittedly, this situation has been improving in the United States and in Canada, but progress is slow, and information about progress is, again, difficult to obtain. The federal statistical agencies in both countries have recently developed some activity in this area. In the United States, we now have access to catalogues from the National Technical Information Service and from the Machine-Readable Archives Division of the National Archives and Records Service<sup>1</sup>. In Canada, we are running behind to some degree. The Machine-Readable Archives Division of the Public Archives of Canada was created in 1974 and has begun to make selected data sets available to researchers. To my knowledge, no catalogue of available data sets has been produced as yet. In addition to the Public Archives, Statistics Canada (the federal statistical agency) has started a Federal Data Clearinghouse, in its Standards Division. This group received Cabinet authorization in November 1974; to my knowledge, no published reports of its activities have appeared thus far. Among the stated objectives of the Federal Data Clearinghouse are the following products:

- a) an information base to enable users of statistical data to locate potentially relevant information sources, related activities and outputs;
- b) a Data Index which will contain, indexed by class of information, variables, indicators and time series appearing in major federal publications and machine-readable data files. This index will reflect the degree and nature of availability of the statistical outputs to various classes of users.

It is not at all clear that the Federal Data Clearinghouse will be of great utility to potential users outside the federal government (except, probably, other levels of government).

It would appear that we need, fairly urgently, some overview of available process-produced data in machine-readable form. The North American Action Group on

<sup>1</sup> Traugott, M. W., and Clubb, J. M., Machine-Readable Data Production by the Federal Government, in: *American Behavioral Scientist*, Vol. 9, No. 4 (1976), pp. 399-401.

Process-Produced Data in the International Association for Social Science Information Service and Technology (IASSIST) has therefore begun to compile an „inventory of inventories“, that is a directory of catalogues which list data bases (primarily, though not exclusively, of the process-produced kind). At this stage, we see the final product of this exercise as an annotated bibliography of inventories to be updated periodically. Although we are nowhere near completion of this undertaking, it is already clear that existing inventories (usually produced by governmental departments) vary tremendously with regard to their informational value to outside users. It would seem fair to expect that any inventory of machine-readable data produced by government provide the following essential items of information on each file (that is, in addition to at least a summary of the physical characteristics of the file, such as number of records, record length, medium in which data are carried, etc.):

1. references to existing documentation, methodological discussions, etc.;
2. an indication of the „degree of availability“ of a file; we would expect that a threefold classification would be sufficient at this first level of information, that is:
  - a. available without restrictions;
  - b. potentially available, depending on circumstances;
  - c: not available under any conditions (e. g. for reasons of national security, or because of restrictive legislation on confidentiality, etc.);
3. an indication of the cost of acquiring a copy of the data set (obviously only under the condition that it were available);
4. the government department; and the name of the position therein, to contact for further information.

Although these would appear to be rather minimal items of information, I should point out that the government of Ontario recently published an inventory of machine-readable data files produced by its various departments in which *none* of the above items were included!

### *B. Information About Specific Data Sets*

Let us assume that we know that a specific data set exists and is, in fact, available to interested researchers. We then require additional information about the data set, for example the following:

**1. The Data-Gathering Process.** Since we are dealing with administrative records of some type, we must assume that the agency responsible for collection of the data (even if, in this case, „data collection“ is no more than some form of bookkeeping or administration) is striving for complete coverage of the event being recorded (I should point out that the same goal of complete coverage exists in the case of enumerations, for example, with censuses. In this regard, process-produced data differ



from sample surveys, where one does *not* aspire to cover the relevant universe. Even within the selected sample, response rates seldom exceed 80 percent.). In other words, when the administrative records describe *events* (for example, bankruptcies, marriages, charges laid by the police, etc.) we must assume that, in principle, all events of this specific nature, which occurred within a given territory and within a particular time-span should be recorded and, as a result, produce an administrative record which describes some pertinent aspects of each event. Again, enumerations are similar in this respect: if the records describe existing elements of population (legal residents, registered voters, institutions of higher learning, mental hospitals, and so on) we must assume that one of the basic aims of the enumeration is to include all members of the population.

In order to assess the degree of completeness of the administrative data set, we must have some fundamental information about the administrative process which yielded the data in the first place. For example: is reporting the event (or responding to enumerations) required by law, or is it a voluntary act? Within the vast array of possibilities among process-produced data, there are obviously wide variations. For example, in the area of crime statistics, the „charges laid by the police“ are frequently an automatic result of the reporting of a crime. That is, the reporting of a crime should be recorded by some police officers; if no record is produced, this could be a result of laziness, inaccuracy, inefficiency or corruption. In other words, the recording of the „event“ in this case is required by law, is carried out by a public servant as a part of his duty, and should have a high correlation with the occurrence of the event itself. In contrast, take the registration of marriages. It is clearly the case that „legal“ marriages (that is, those which are accompanied by some formal ceremony) are normally registered, when the registration is part of the formal ceremony itself — as in most civil marriages. However, if the formal ceremony is a religious one, taking place in a secular nation-state, recording of the event may or may not take place; such recordings may not have any relevance with regard to the *public* recording of marriages. To make matters even more complicated, common-law unions clearly are not registered by any administrative process. Thus, civil marriages *must* be registered, since such registration is a legal requirement. However, as a measure of the „nuptiality“ of a society, such records may be quite inadequate, since in most societies one is not legally required to register the beginning of some period of heterosexual cohabitation.

In cases where registration is required by law, we should know that the sanctions are in the case of non-compliance. (It should be obvious that the sanctions are different in the cases of voter registration and filing an income tax return.) Other questions which fall in this domain are: to what degree are laws regarding compliance in fact enforced? To whom, or which agency, does one report? How much time is permitted to lapse between the occurrence of an event and its registration? Are the events located in *time* (time of occurrence of the event, some measure of duration of the event, some measure of duration of the „event“ or both) and *space*?

To make matters even more complicated in this area, let us consider yet another aspect. So far I have assumed that some event occurs and that *someone* is expected

to register the event, that is, to produce some administrative record. I have assumed that the main source of discrepancy would be the non-existence of the administrative record. But there is, of course, a second type of discrepancy; this occurs when an administrative record exists which has *no* counterpart in the universe of events (for example, to file a claim for insurance benefits or some form of public welfare, based on some fictitious event).

It is my opinion that these kinds of information are generally quite difficult to obtain. Often, documentation of this nature exists only in unpublished „working papers“ and memoranda, for which bibliographic entries are scarce or non-existent. Moreover, it is not always clear *who* (or which branch of a particular agency) is responsible for the external distribution of such documents. Not uncommonly, documents of this nature are labelled as „classified“ and are thus totally inaccessible to external users.

I have tried to develop a fairly lengthy list of questions in this specific area, because it is especially in this regard that process-produced data differ from most (if not all) other types of social science data. It is my impression that a large proportion of the researchers using these data is insufficiently critical of the reliability and validity of available data and has failed to ask pertinent questions which could reveal the quality of the data.

**2. Administrative Decisions Regarding Data Manipulation.** Although we are working from a definition of process-produced data which specifies that they are by-products of administrative processes, it is obvious that virtually no administrative agency can handle such records in their original state. As all other machine-readable data (I am obviously, in this section, not terribly concerned with process-produced data which only exist in manuscript form), such records must have gone through processes of recording, transcription, coding, keypunching, reformatting, editing and recoding. It is inconceivable that all these procedures are foolproof. Thus, errors slip into the records at various phases of the manipulative phase. As a result, it is likely that, at various points, decisions have to be made regarding any errors or inconsistencies which can be detected. Such documents as coding schemes, ways in which missing values were handled, procedures for editing and cleaning data are all essential for the proper interpretation of a given data set by a secondary user.

### *C. Governments as Research Resources*

The two sections above reflect a broader dimension, which also needs to be discussed. It is quite clear that governments, at various levels (i. e. national, regional, local) have a virtual monopoly over a large share of the existing process-produced data. This is most obviously (and trivially) the case when we deal with records which are the by-products of internal governmental processes (e. g. minutes of Cabinet meetings, federal budgets, and so on), but it also extends to data which are

by-products of administrative processes relating to other sectors of society. In part, of course, this reflects the expense involved in gathering data on a massive scale (for example to record every birth occurring in the country); in part it reflects the fact that governments have legal, political, economic and persuasive powers which private researchers can not bring to bear.

The fact that information about process-produced data is either non-existent or relatively inaccessible may well be taken as an indication that the relations between the social science research community and governments, as data collectors and sources of data, have not yet been worked out satisfactorily — at least in North America. It is my impression that lines of communication are quite imperfect. The research community appears to have relatively little impact on the collection, manipulation or distribution of data by various governments; prospects for improvement of this situation do not seem to be very good<sup>2</sup>.

In general, social scientists appear to have had some success in acquiring census data in machine-readable form, both at the aggregated level and at the level of individual records. The United States quite clearly took the lead when it released public use samples of the 1960 Population Census. This initiative was followed in 1970 by an even more generous program of access to individual data. Canada was a somewhat hesitant follower with its set of public use sample tapes from the 1971 census. As the recent article by Judith Rowe<sup>3</sup> indicates, several other countries are prepared to give researcher access to at least samples of individual census records, after appropriate safeguards against disclosure of identification have been taken. It is indicative of the state of affairs with regards to administrative records that no equivalent programs exist to give researchers access to individual records in this area. Three causes can be mentioned for this state of affairs: first, laws regarding access to information are often unduly restrictive, often making access to data impossible for individual researchers; secondly, even when the legislation allows for some discretion on the part of governmental authorities, the importance of the data for research is frequently not recognized; and thirdly, the research community has generally not produced the kinds of efforts regarding administrative records which eventually yielded the various public use samples of population censuses.

It would seem necessary that the social science community increase, where possible, its involvement in the shaping of further legislation on freedom of information, privacy and access to governmental data. Moreover, strategies should be developed to facilitate the flow of data and information within the constraints set by existing legislation. For example, it may be useful for someone to produce an unbiased discussion of the various techniques by which records can be protected against disclosure of identity (such as collapsing of sensitive categories in some variables; elimination of unique or near-unique identifiers, introducing random disturbances into

<sup>2</sup> Op. cit. pp. 402–407.

<sup>3</sup> Rowe, J., *Non-American Census Data in Machine-Readable Form*, in: *Demography*, Vol. 14, No. 1 (February 1977).

some proportion of the records and some proportion of the variables; scrambling groups of records, and so on). Finally, there appears to be a great need to educate various „publics“ (governmental decision-makers, politicians, voters, the general public) about the need for access to public data.

### III. Problems of Documentation

Let us assume that we have acquired a hypothetical process-produced data set, accompanied by the types of documentation outlined so far. We will then still need several items of documentation which may, or may not be included.

#### *A. Specification of the Universe*

To use the terminology employed in the description of sample surveys: we need to know how the relevant population is defined, or which members of the population are defined as being part of the data set. (Remember that I am referring to a population of *events*, not necessarily a population of individuals.) To indicate the effects of different universe specifications on the nature of the data produced, let me describe the Canadian data on unemployment. There are three sources of information: the Monthly Labour Force Survey, Statistics from the Unemployment Insurance Commission, and the population census (Note that only the second source is of the process-produced variety). The Monthly Labour Force Survey is „... a multi-stage probability sample of the civilian, noninstitutional population of age 14 and over of Canada excluding the Northwest Territories and the Yukon“. The census, on the other hand, covers the total Canadian population<sup>4</sup>. Thus, residents of the Northwest Territories and the Yukon are included in the census (but not in the Monthly Labour Force Survey). Moreover, the census includes the institutional population, the Armed Forces, Indians on reserves and Canadians resident abroad. On the other hand, the census questions on employment status are only asked for persons 15 years and over, while the Monthly Labour Force Survey has a lower age limit of 14.

While the former two data sources refer to the *prevalence* of unemployment, statistics from the Unemployment Insurance Commission refer to the *incidence* of employment, among a specific population. The statistics refer to the claims for un-

<sup>4</sup> Dominion Bureau of Statistics, Canadian Labour Force Survey (Methodology), Ottawa 1966, p. 8.

employment insurance benefits (thus, claims are the *events* of which a record is kept). Therefore, if we assume that each claim represents one and only one claimant, the statistics will give us the number of claimants of unemployment insurance benefits. Claimants must, by necessity, belong to the insured population. This population is *not* identical to the total civilian noninstitutional population. It consists „ . . . mainly of the paid worker segment of the labour force. Members of the Armed Forces are in the insured population“<sup>5</sup>. There are, however, exceptions. For example, employed persons with very low earnings, employed persons 70 years and over, and employed persons to whom a retirement pension under the Canada Pension Plan has at any time become payable do *not* form part of the insured population (Statistics Canada, 1973: 81–82). To make life even more complicated, the census definitions of „employed“ and „unemployed“ are not identical to those used in the Unemployment Insurance Commission's statistics.

I have spent some time delineating these differences in order to illustrate the importance of knowing the *exact* definition of the universe to which the data pertain. It is not always easy to find the exact specifications as they were stated in principle. It is often even harder to find out how membership in the universe was determined in practice. The preceding discussion, by the way, also illustrates one of my earlier points: the difficulty of obtaining this type of information about existing data sets. All the above information about variations in universe specification can be found in an unpublished discussion paper from the Labour Force Survey Section of Statistics Canada. Such papers do not normally appear in the common bibliographic tools (such as library card catalogues, and so on).

### *B. Estimates of Errors of Coverage*

Demographers have traditionally made a distinction between errors of coverage and errors of content. Errors of coverage deal with violations of the assumption that members of the universe are presented by a record, once and only once. That is, the assumption that there is a one-to-one relation between an event (a birth, death, bankruptcy, admission to a mental hospital, and so on) and a record in some administrative data file. Thus, errors of coverage are of two kinds:

- a. the exclusion, or omission, of elements which should have been included, and
- b. the inclusion of elements which should have been excluded.

In the former category, we can distinguish the following subtypes:

- an event occurs which is not recorded (e. g. failure to register a birth, death, etc.);
- an event occurs, is registered but (incorrectly) defined as not belonging to the particular universe;

<sup>5</sup> Statistics Canada, Concepts and Definitions used in the Canadian Labour Force Survey, Ottawa: Labour Force Survey Discussion paper, p. 8.

- an event occurs and is registered, but the record is „lost“ during the phases of recording and data manipulation.

In the latter category, possible subtypes are:

- a registration is made for an event which did not really occur;
- a registration is made for an event which did occur, but which did not belong to the particular universe;
- an event, which did occur and belonged to the particular universe, was recorded more than once;
- during the phases of recording and data manipulation, multiple records of a single event were erroneously produced.

For any given data set, we obviously require not only a discussion of the possible sources of error, but also some estimates of the magnitude of the various types of error. Unfortunately, we usually find ourselves in the position where no such estimates exist, or where they are severely biased.

### *C. Estimates of Errors of Content*

The second kind of error is usually called „errors of content“. In this case, an event has been recorded, but at least one of its characteristics has been recorded incorrectly.

Of special importance in this respect are those variables which anchor the event in space and time. Let me, again, give a few illustrations from the vital statistics field.

Temporal measures may be in error in several ways. For example, births may be recorded with an indication of the date of birth, or of the date of baptism (as happened with many early birth records). Quite clearly, under conditions of severe infant mortality, the two points in time would lead to drastic differences in estimated birth rates (because infant deaths which occurred prior to the baptism would, of course, not require a baptism. In such case, it is likely that no registration occurred at all). Another problem can occur when an event takes place close to the end of a calendar year (or some other unit of time over which information is summarized). In such cases, the registration of the event will often take place in the following year. This may lead, trivially, to an error in recording the date of occurrence. But another possibility arises. Usually registrations of vital events have a „cut-off date“. For example, if an effective period of three months is allowed to pass between the occurrence of an event and its registration, one can „close the books“ for calendar year  $y$  on April of year  $y+1$ . Now suppose we have an event which occurred at some time during year  $y$ , but which was not registered before the cut-off date in year  $y+1$ . Among the possibilities are: the event could show up in the summaries for year  $y+1$ ; it could, later, be included in revised statistics for year  $y$ ; it could be omitted from summary statistics altogether. I should indicate that this last possibility is by

no means just the figment of someone's fertile imagination: during the 1960's, an estimated 10,000 births per annum were indeed, in this fashion, excluded from the vital statistics of the province of Quebec.

Such problems are obviously of great importance when one studies variations of phenomena over time. In a similar fashion, spatial measures are of special importance. Each event which is recorded will have at least one spatial referent. In some cases, such delimiters are of no value (for example, data on attendance in the House of Commons), but those are exceptions. Normally, we will require clear specifications of which measure is recorded. For example, vital statistics require the recording of either the place of usual or legal residence (of the mother, in the case of births) or the place where the event occurred, or both. In the case of births, it is not uncommon for residents of small towns or rural areas to have their babies in nearby cities. If we were using data based on „place of occurrence“ instead of the usual residence of the mother, we may find artificially high birth rates for cities with maternity wards in hospitals, and artificially low rates for places without such facilities. I should point out that this type of distortion may indeed have led to some erroneous statements regarding ecological variations in the incidence of mental illness (derived from statistics on admissions to mental hospitals).

In addition to these „anchoring“ measures, there are, of course, other variables which may contain errors. We need to know, in fairly great detail, what experience the administrators have had with these data: which types of error are most common, what procedures have been followed to isolate errors, how errors have been corrected, and which methods have been employed to validate the corrections.

#### *D. Codebook Equivalents*

We should distinguish two kinds of process-produced data: they either refer to single events, or they are the products of various computational procedures and take the form of aggregated data, or summary statistics of some kind. In the former case, we are essentially dealing with codebooks which differ little from those which accompany sample surveys. We need to identify variables, give the exact wording of the questions, indicate the various permissible response categories, and so on. Ideally, we would also like to have a sample copy of the registration form. Aside from the fact that relatively few data sets are available in this format, and that in most cases nothing like a codebook exists, we are not faced with unique problems.

In the latter case, where there is no longer a one-to-one relationship between the machine-readable record which the researcher gets and the originating event, problems do arise. As an example take the simple case where a data file gives us total populations, at a specified point in time, for areal units in a country. Quite clearly, we will require a statement about the universe covered (Note that this may be a somewhat different universe from the one used to collect the data. For example, many

countries do not provide published data on the armed forces). In addition, one would like to be given summary statistics on such variables, such as minima, maxima, means or medians and standard deviations.

Finally, in the case where the data refer to elements which have spatial counterparts in the real world, a codebook should have sufficient information to locate these elements on maps (or in the real world). One possibility might be to include, with the data, coordinates of sufficient detail for one of the common computer mapping packages, so that users may produce their own maps.

## IV. Problems in Data Transformation and Analysis

Although I am now moving beyond a general interpretation of „access“, let me briefly enumerate the following issues:

### *A. „Ragged“ File Structure*

Often process-produced data have a file structure known as „ragged“: logical records do not have a fixed length, but their length varies in relation to the number of elements they contain. For example, data pertaining to households could have formats varying with the number of persons in the household. Most of the standard computing packages do not lend themselves easily to the handling of such data files.

### *B. Simultaneous Use of Several Records within a File*

Most statistical packages do not allow the simultaneous handling of variables from more than one logical record. Both in the analysis of time-series data, and in an analysis of contiguity, one runs into difficulty with the most common packages.

### *C. Peculiarities of Aggregated Data*

Most statistical packages appear to be the products of the research style of survey researchers. As such, techniques which are more suitable for the analysis of aggregate data tend to be rather poorly developed.



#### *D. Problems of Data Linkage*

Almost invariably, the analysis of process-produced data relating to *events* requires the use of the corresponding enumeration (to standardize for variations in the exposure to risk, for example). Thus, *two* data sets are often involved. The following problems may occur:

**1. Standardization of stimuli:** in its most fundamental form this problem manifests itself through differences in definitions of the universe. For example, one would have problems if one were to use the Canadian unemployment data from the Unemployment Insurance Commission in conjunction with data from the population census to calculate measures of the incidence of unemployment in Canada.

At a less fundamental level, we may find cases where the universes do coincide, but where characteristics are measured in slightly different fashions. This problem is a common condition when the „linkage“ involves data from two points in time. For example, the Canadian census definition of „mother tongue“ changed between 1931 and 1941. The earlier criterion required the language still to be spoken, while the later criterion only required the language still to be understood.

**2. Standardization of responses:** even in cases where the stimuli are identical, the available sets of responses may not be exactly identical, either in terms of available categories or in terms of the operational definitions of some of the categories. For example, anyone studying illegitimate fertility in Canada has to come to grips with the fact that, since 1949, Ontario has adopted a different definition than the other provinces: in the former case, births are illegitimate if the mother was never married, while in the other provinces births are illegitimate if the parents are not married to each other.

**3. Standardization of delimiters:** as I already pointed out, variables dealing with space and time are of special importance. In North American society, we are faced with many ways to divide territory: municipalities, counties, planning areas, school districts, electoral districts and, in Canada, the „bilingual district“. For reasons which are totally mysterious, the boundaries of such sets of districts usually overlap. Moreover, boundaries often change over time. Therefore, any „linkage“ of aggregated data sets is likely to present the user with areal units which are not perfectly identical and which can not easily be mapped onto each other. We clearly need estimation techniques to deal with such situations, since one can usually not gain access to data at such level of areal detail that a reconciliation of boundaries can be established.

## V. Conclusion

I have attempted to provide a list of some problems one is likely to encounter during the various phases of accessing and handling process-produced data. I have tried to indicate that most of these problems are nearly unique to process-produced data; thus, we gain little from the experiences of survey analysts. Given the enormous variety of process-produced data which now is beginning to find its way into the research community, solutions for the array of problems indicated are likely to come from persons working in different scientific disciplines and in different countries. Any efforts to increase and facilitate communication between these highly dispersed researchers should therefore be strongly supported.

## A Perspective on Social Science Data Management

This essay makes the following arguments:

1. The available social science „software packages“ only service two of a dozen data formats needed for social science research; the lack of appropriate software has inhibited social science development.
2. The large scale efforts on „database management systems“ (DBMS) including the work of the CODASYL Programming Language Committee, are only marginally relevant and do not address the bulk of social science needs.
3. The social sciences would be greatly aided by an effort to delineate their needs more clearly and spearhead appropriate developments.

*Current software.* Most social scientists think of statistical software in terms of the statistical tests and processing options offered. Does a package offer repeated measures analysis of variance? What are its missing data options? Questions like these are usually uppermost in a reader's mind when scanning a new manual for a software package.

Another issue, however, revolves around what kinds of data formats the statistical package can process. While most statistical packages offer procedures for adding a new statistical test (by calling a user-written FORTRAN routine or the like), they are quite unable to accommodate new kinds of data structures.

Veterans of social science data processing will immediately associate the „data format“ problem with the myriad of ways that numbers tended to be represented in early computers. The habit of punching multiple responses in one card column for counter-sorter tabulations led to the necessity of binary column options in early software. Such software also had to accommodate to different kinds of keypunch and paper tape formats, with basic incompatibilities between the magnetic tape formats of different manufacturers. As the representation of numbers became more standardized, new problems arose in the standardization of procedure for labeling data storage volumes (including disks and tapes) and labeling the files stored on those volumes.

The „data format“ problems considered here assume that issues of how the numbers are represented and how the files are labeled are solved (even though such standards seem to be frustratingly slow in coming). It further assumes that each data file can be described by a „codebook file“ that tells the user what each piece

of information stands for and tells the statistical package how it is stored in the file. The user can thus refer to each variable by name. If the data base manager decides to change the form in which the information is represented (e. g. a switch from decimal to binary), the change can simply be entered in the codebook file and remain invisible to the user. (However, the social science user is usually not as „independent“ from data specifications as data base specialists might suggest. For example, the social scientist needs to know the precision with which quantitative information is stored in planning statistical analyses.)

The „data format“ issues treated here instead focus on the fact that social science data need not come in a neat organization that can be efficiently represented by a „rectangular array“. A „rectangular array“ is one with a fixed number of rows and columns. „Efficient use“ means that many of the cells are not empty or repeating information elsewhere in the array.

Existing social science software services two kinds of rectangular data representation. In one format („cross-sectional“) the rows are respondents (such as in a survey research questionnaire) and the columns are the responses to the questions asked. Both the rows and columns are unordered. The respondents may be in the order in which they were surveyed, but this has no bearing on the statistical analysis. In one survey, the respondent's age may be in columns 22–23 and a code for the birthplace in columns 49–50 while in another survey these two variables may be stored elsewhere. The codebook file may directly inform the software that respondents' ages are recorded in columns 22–23, thus making the location no longer a user concern.

A second kind of social science rectangular format, associated with econometrics, has variables represented in different rows and points of time represented by the columns. The variables are unordered, like the variables in the previous format. However, the columns are ordered, with successive columns representing successive years, quarterly periods, months, or whatever is the unit of time. A variety of time series procedures can be carried out on the ordered dimension; for example, one time period measure can be lagged against another. If data for a cell is missing, one may interpolate an estimate from the adjacent time period information.

These two kinds of rectangular formats have other differences. Often the survey rectangular array may be much larger than an econometrics rectangular array. A cross-sectional study may have 500 pieces of information on 20 000 people, or an array of 10 000 000 cells. An econometric data base may have 500 measures on 200 time periods, or 100 000 cells. The difference is more a matter of scale than logic, but it has logical implications for what kinds of data storage strategies are appropriate. When the National Bureau of Economics Research adapted their time series package (called TROLL) to cross-sectional data, it turned out that the cross-sectional facility, while offering attractive processing features, was not suitable for large files. (Conversely, SPSS, the most used statistical package for cross-sectional formats, now includes a „lag“ option, but this is quite primitive compared to the intensive analytic offerings of econometric packages.)

There are two major ways in which the rectangular arrays are processed. One is

to pass all the data through the computer in order to compute the answers. All the information in one row (that is, all the information for one subject) is read in and the information in the columns needed for the analysis is selected out for statistical computations. This process is the usual strategy used by Data-Text, SPSS, OSIRIS and other well known cross-sectional format packages. The second strategy is to have the user declare what variables are going to be used beforehand and select them out for storage in computer memory or a high-speed access device. Thus, in any statistical analysis, only the subset of variables specified beforehand are passed through the computer. For cross-sectional formats, this is the strategy used by IMPRESS, the Dartmouth interactive system. A survey may have 500 variables, but the user on a particular investigation may only have need for 10 of these for an analysis. The prior selection thus results in a 50:1 reduction in the amount of information passed through the computer. This second strategy is used by most packages for analysis of time series formats. An econometric data base may have hundreds of variables stored in it, but an analysis may only require some five or ten of them at one time. By making a prior declaration, the computer brings the variables from background storage up to foreground, for repeated examination in the analysis.

*Other data formats.* Having said that the social science software packages service two major formats, we will now propose that the social sciences need to have facilities for considering ten other kinds of format. These additional ten formats are only for non-textual data. If one wants to include formats for textual data (a topic dear to this writer's heart) the number of formats greatly expands.

In 1975, the *SPSS Newsletter* asked readers to tell what kinds of data applications they had for which SPSS was inadequate. A couple of hundred responses were received and the folks at SPSS Inc. kindly provided a xerox of the replies. What were the major missing capabilities perceived by the users?

By far, the most cited one was the inability to handle hierarchical files and the most cited application of this was to census information. If a file, for example, is organized by region, town, tract, street, block, building, family unit, and persons, it is extremely wasteful to repeat all of the above information for each person. A hierarchy of levels means that all of the information at a higher lever is applicable to each of the units at lower levels until the next time information at the higher level is changed.

A hierarchical file can be organized in several ways. If all the analyses are to be at the lowest level (e. g. the person), then one file may be created, with each record identified as to its level. The package would use the higher information to fill out a profile and process a unit each time a record is encountered at the lowest level. The profile for a person would be the person record, plus all the current higher levels previously encountered. If, however, many of the analyses were to be at the level of the family, or the level of the housing structure, then it may be efficient to have the higher level information on separate files. A housing study can then be processed

without going through all the information about persons. Each record needs a marker indicating when it is the last record at that level for a unit. Thus, the record for the last family member may have a marker indicating the next family record should be read while the last family in a building may have marker indicating that the next building record should be read, etc. Although it is not difficult to prepare procedures to read such hierarchical information, the lack of standardization has made social science software suppliers hesitant to invade this domain. The U.S. census bureau has such hierarchical procedures, but they are not common to most cross-sectional software packages, except in some very rudimentary ordering of major and minor files.

A second common inadequacy (although much less cited in the SPSS survey than the first) concerns situations where there is a string of data. For example, a multiple response to a question such as „what do you like about the current presidential candidate“ may produce a few replies from some respondents and many replies from others. Most existing packages allow the respondent to set up a fixed number of columns (corresponding to the *most* replies received) and to say a response occurs if it is coded for any of these columns. Often this is done by an „or“ specification in which the user says that the response exists if it is coded in columns X or Y or Z, etc. Some packages such as CROSSTABS allow for a string option where the response is recorded if a code occurs in string S, where string S has previously been defined as a list of columns. The string option thus saves repetition of all of the „or“ specifications.

There are two problems with this rudimentary string capability. One is that the storage is wasteful if the short strings representing the few replies of most subjects have to be dummied out until they are the length of the longest response string. If numerous multiple response questions appear in any one survey, then the normalization of the data into a rectangular array may involve numerous blank cells.

The second problem is that the packages do not provide ways of handling order within the string. Let us say that the string is a job history, with each unit in the string one of a series of jobs, and the jobs are ordered in terms of recency. A question regarding how many people held such and such a job before they held another kind of job is difficult to search with most packages. People responding to the SPSS Newsletter cited a variety of such history files, including school history, police record history, credit record history, medical history, geographical and housing history (one record for each move made), etc. In each case, the data is ordered by time, but there are a varying number of units per respondent.

In contrast to our neat rectangular array, we call this kind of data „ragged“. Unless we normalize with blanks into a large rectangle, there will be a varying amount of data for each subject. Each subject may have one or more strings associated with the record and each string may be of varying length, producing a varying length record overall.

Occasionally, there is a situation where there is but one long string for each respondent, but it dominates the record. One example is the time-budget study, where each record consists of some basic information about the respondent's age,

sex, family status, work status, education, housing, etc., followed by a string where each unit is an „event“ and the events cover a span of an hour, a day, a week, or whatever. In this case, one may wish to consider this a form of hierarchy in which the unit is the event rather than the person, and the basic information about each person is higher order information. The ordering from one event to the next can be examined by a „lag“ feature from one unit record to the next. A variety of other chronicles also take this form, such as that of a small group meeting.

If the bottom unit in a hierarchy is an event, then the event may take several different forms, with data for different events being different sizes. For example, if the next thing a person does is watch television, we may want to know what kind of program is watched, and who watched the program with the respondent, while if the next event is a telephone conversation, we may want to know other kinds of information. Each bottom unit could then be coded to inform the package which kind of end unit it is.

In other instances, investigators have preferred to score life histories so that medical history is separate from educational history, yet there should be ties so one is able to piece out time overlaps and time sequences. The separate histories simplify applications for secondary analysis, where an investigator interested in medical history need not pass through educational information unless it is wanted. In these instances, software is totally inadequate and life history projects in United States, Norway, and Poland, among others, have had to struggle with their own software. It might be thought that the coordination of different kinds of history strings for one respondent might be especially related to current work in data base management systems.

Partial orderings may be partitioned in larger „chunks“. In a life history, we have a number of events with at least several kinds of information (what, when, where, with whom) for each event. In other kinds of research, we have fewer „events“, with more information about each one. A panel survey, in which the same people are interviewed at different points in time, can have each survey regarded as one event. In experimental research, data from each subject may be gathered at different time periods in the course of the experiment, with each repeated measure analogous to an „event“.

A variety of further compounds, however, may occur. For example, the Institute for Social Research at The University of Michigan has recently employed a panel design to gather time budgets for the four seasons of the year. Each survey includes both the husband and wife in the same family. Thus, the data consists of eight varying length strings (the time budgets), each paired (husband and wife on the same day) at one sample day for each of the four seasons of the year. One can imagine analyses made of this data set that would involve searching across strings of couples at different times of the year for comparisons.

Another form of chunking is where groupings are made but the groupings themselves are unordered. A cross-national comparative design may have the subjects grouped by country. In experimental work, subjects may be grouped by treatment. Grouping of subjects may be combined with ordering of variables. A cross-national

study of time budgets, for example, may group subjects by country. An experimental design may have repeated measures for each subject, with subjects grouped by treatment. Such groupings may be considered a rudimentary form of hierarchy, with the subjects in each group implicitly sharing common characteristics, even if it is no more than a code indicating treatment type.

At this point, we have expanded our perspective from:

*current software:*

- a) rectangular, unordered both dimensions,
  - b) rectangular, ordered on one dimension, unordered on other,
- to various combinations of

*needed software:*

- |                |                      |   |
|----------------|----------------------|---|
| a) rectangular | c) ungrouped         | f) unordered  |
| b) ragged      | d) grouped           | g) ordered on one dimension                                   |
|                | e) grouped hierarchy | h) partially ordered (partitioned chunks or embedded strings) |

*Order versus ordering.* In most data analysis applications, any order on the data is from time sequences. Another aspect of order is the rearranging of data so as to reveal patterns.

A mechanical ordering procedure was once used to produce Guttman scalings. As described in *The American Soldier*, a box was created with slats, in which each slat represented the responses of the subject, and the columns were arranged so as to give first priority to the item getting the most yeas, the next most yeas, etc. The slats were then rearranged so as to group most of the yea marks on one corner of the rectangle.

Modern computer techniques can be used to rearrange both whole rows and whole columns at a time so as to make order apparent where there was little to be seen. If all responses are either zero or one, the ordering can be either to maximize areas of zeros or maximize areas of ones. H. White, R. Breiger and associates have developed techniques for rearranging matrices so as to group clusters. Suppose the rows are persons citing and the columns are persons cited. Initially, they may be unordered. If they are rearranged to form clusters, we may find (as, for example, Breiger does for journal citations) a group that cites each other, a set of asymmetrical intersections in which the group cited does not reciprocate, and a set of intersections where people do not cite each other.

Thus, we should distinguish data sets having order over time from the mechanical ordering process to uncover underlying relationships. The ordering process links file handling to measurement theory. One advantage of the mechanical ordering is that the unexpected case, that is the case that is the case that appears where not expected once ordering takes place, is readily salient.



*Graphs and list structures.* If a group is relatively small, then the combinations of persons choosing and persons chosen can easily be represented by a rectangle. A larger information structure may not lend itself to such rectangular convenience. For example, if the membership of all corporation boards are listed, one may wish to produce an index of interlock between different boards, and cluster the corporations in terms of the connectedness of their board membership. This structure again requires different management software, such as represented by the „Baron“ program by Levine for representing interlocks.

A list processing system represents data as a series of lists. A list may be composed of items, much like a grocery list, or it may also contain the names of other lists. The system can answer questions like: Is X on list A or any list named by list A? Which lists contain X? Are X and Y on a common list or any lists connected by a higher order list? How many such connections exist between X and Y in the entire data domain?

A central feature of list processing languages is the use of pointers. Each item in a list, in addition to containing a pointer to the contents of that item, also contains a pointer to the next item on the list. Once one enters a list, one can use the pointers to find one's way to successive items on the list. If a list is symmetrical, then each item contains pointers both to the previous item, as well as the next item. Thus, one can find one's way back up the list as well. In terms of the computer, a pointer is nothing more than an offset address, relative to the workspace available for list processing.

The development of list processing and graph procedures has been more extensive in cognitive information processing, especially as represented at Carnegie Mellon University by Simon, Newell, and others. Its application in sociology has been rare and in some cases it was used where it was not necessary. List languages have their costs, as do most data base management systems, so one must generally consider what features are indeed necessary. If a rectangle or a simple hierarchy will suffice, then it is by all means to be preferred, especially if the data base is large.

*Summary:* The various capabilities we have cited may form odd combinations and they do not summarize in a simple table. However, some of the most important thus are as follows:

— Rectangular

- 1) — both dimensions unordered
- 2) — one dimension ordered
- 3) — one (or more) dimensions grouped
- 4) — ordering on two or more dimensions

— Hierarchical

- 5) — units unordered
- 6) — units at one or more levels ordered
- 7) — varying end units

- Ragged (varying size units)
  - 8) — single string for each unit
  - 9) — multiple embedded strings
  - 10) — ordered versus unordered strings
- Networks or graphs
  - 11) — structures
  - 12) — processes

Both hierarchies and ragged data forms may logically be an economy of storage for information that can be normalized into a rectangular array by data repetition or padding with blanks. Both ordered data, and the capacity to order data, involves different data referencing and data processing capabilities than packages that limit themselves to unordered data. Networking or graph procedures get into elaborate (and time consuming) pointer systems that may be necessary for some data forms where a matrix would be extremely sparse. At present, the available social science packages focus on the first two capabilities in the summary list.

*Role of Data Base Management.* Much of the work on commercial data base systems is oriented towards inventory or client list problems. One may have a large inventory of parts and want to order it both in terms of supplier, warehouse location, and object use. Clients may be ordered by location, salesman, speciality, etc. If there is considerable repetition, then one may substitute a code for a complete entry and use a table lookup to get the full entry, or one may turn to a hierarchy so that entries are grouped by their sharing of common information.

If, however, a hierarchy is used, then it is difficult to access information that crosscuts the hierarchy. If parts are organized by object use, then it is difficult to retrieve those that come from the same location other than by searching the entire file. Much of data base management is concerned with threading various indexing schemes through the same data base, either by the use of pointers in relational arrays, hierarchies, or networks. Major emphasis is given to designing such structures to facilitate adding, deleting, updating, and retrieving.

Most data base management systems are for files that are continuously undergoing change, like a current inventory and client list, so that there is a willingness to pay a price for abilities to continually make updates and other changes. In contrast, most social science problems are more set. One may think of a social indicator project, for example, as a matter of repeated updating, such as National Opinion Research Center annual General Survey. But this is updating of a different kind. Once the 1974 survey is completed and edited, for example, it remains constant. The 1975 and 1976 additions will be tacked on as additional units in the data base, but the 1974 unit will not be subject to repeated change. If the social indicator is based

on panel techniques, a simple respondent identifier may tie one unit to the next. Past units then are relatively fixed; not to be molested unless a coding error is discovered.

With little emphasis on updating, social science data entry and storage problems are less difficult. Data entry, both for raw data and codebooks, is straightforward, with routine entry and editing procedures being quite satisfactory. Once the data is checked, it can then be made more compact for final storage, including the use of bit strings and binary representations instead of digital storage. The documentation of this compacting is made part of the machine readable codebook accompanying each data set.

Although we have emphasized the range of social science data formats, the bulk of day to day processing, of course, will continue to be in rectangular format and handled by conventional software. Two level hierarchies may well be handled by social science software such as the Norwegian DDPP system; the „group“ option within this software allows for higher level information (e. g. households) to be separately stored, but automatically cross-referenced, with the lower level data (e. g. respondents). Existing data base management systems, such as the IBM *information Management System's* hierarchical approach built to CODASYL specifications, may well be useful for some social science applications. More recent developments of relational file management strategies, using the data management algebra and calculus developed by E. F. Codd, may also prove useful.

Before adapting any system, the user does well to examine the speed and capabilities of the system, making certain it is neither unnecessarily costly or unnecessarily constraining. So far, the experience with attempting to use commercial data base management packages on complex social science data has been frustrating, with users often ending up writing their own programs. This situation is especially true where data comes in many forms and often involve linking varying levels and intensities of reporting. Large scale evaluation research projects, including the areas of negative income tax, health care delivery, and housing programs, are particularly notable for such problems.

One system that may prove useful for a wide range of applications is the SAS statistical package, developed by the SAS Institute in Raleigh, North Caroline. This system separates data statements from procedure statements. The data section employs a pointer system, with the ability to read data in either a stream or user edited mode. The data section can contain GO-TO transfers so as to provide the flexibility of a programming language. The same input file may be referenced in different statements, each time advancing the pointer (or backing it up) by a controlled amount. The data section may be used to combine selected data from several input files; its features include extensive merging options. Data records may be of fixed or variable length.

The result of the data section of SAS is an intermediate file, much like the intermediate files of SPSS, DATA-TEXT or OSIRIS. In fact, SAS can read the intermediate files of these other software systems. The statistical procedure section may be then called, offering a wide range of advanced statistical procedures, plus the op-

tion to access the BMDP statistical package subroutines. Recording and regrouping is accomplished by „sort“ or „matrix“ procedures.

In future work, the fundamental distinction between data handling and statistical analysis procedure stages may be more carefully observed. Although we have emphasized the hierarchical, ragged character of many kinds of social science data, many advanced statistical analyses require cleaned up rectangles (such as correlation matrices) or triangles (such as symmetrical distance measures) as input. The statistical procedure section should take care to hold intermediate matrices (such as cross products) so as to not have to recalculate them each time a new statistic is used.

We may envision various data base management systems as serving the data stage, well interfaced to turn data over to analysis procedures. A suitable overseer procedure may be needed to supervise the passage of control, or it may just be that the systems „know about“ each other. This knowledge might seem only natural; however, there has now been more than a decade of fuzziness, with the most used statistical software systems not being able to read each other's intermediate files.

Is a fundamental separation of stages, with many interfaces, appropriate to a „user oriented“ software? I think so, providing it is within reasonable limits. Special care is needed that the data management side is both flexible and natural to the user. In my opinion, an extended APL (to handle trees and other nonrectangular structures) may be natural to a minority of social scientists, but its economy of notation does not come naturally to the community at large. On the other extreme, a COBOL type of pseudo-English is not needed either. Again, SAS has offered a suitable, but highly flexible, intermediate.

Several examples help to point out what has user acceptance. The IBM statistical subroutines, for example, do not find high popularity among social scientists. Part of the reason is that they lack many needed built in features, such as missing data options. The other reason is that they assume that the users write much of their data handling programming. Similarly, „coherent programming“, which allowed the user flexible power in putting together building blocks, again was too complex to win large audiences. On the other hand, a simple language like *Utility Coder* offer remarkable power in data handling, and is easily interfaced to a statistical analysis mate such as *Crosstabs*.

The level of effort in social science software programming remains small compared to many commercial centers. Just one division of the American Telephone Company, New England Telephone, has over 400 programmers. Some companies probably have a larger level of effort than all university based social science programming put together. The social sciences can no longer accept the lore of how a few faculty and assistants built systems in the 1950s. As Brook's book, *The Mythical Man-Month*, well describes, this lore tends to lead to mythical estimates and even the belief by old timers, that they can do it again in today's settings. As Brooks shows, the costs of coordinating and documenting increasingly complex systems multiplies almost exponentially. With limited financial resources, the social sciences thus do well to watch the commercial ventures closely, as well as developments by government agencies such as Census and Social Security.

In order that social scientists can understand their own needs better, we need to go beyond the SPSS user survey and carefully examine the different kinds of analysis problems that arise. How then can the spectrum of problems be best partitioned? Which tasks are adequately covered and where is more development most urgently needed? In some cases, such partitioning may bring together researchers from quite different fields who were previously unaware that they had data analysis formats in common.

Once the many disciplines of the social sciences, including economics, demography, geography, sociology, history, psychology, political science, evaluation research, education, survey research, group dynamics, and anthropology have been integrated into a suitable analysis framework, our future steps may reflect considerably more wisdom. By identifying commonalities, it may be that enough different groups show a shared data handling problem so that a larger systems endeavor is more appropriate than any one group could justify. The commonalities may help some groups understand better what their data problems are and the range of options open to them. A reference point can be established for future planning.

## VIII. Preservation, Storage, and Access



## Problems and Procedures for Preservation and Dissemination of Computer-Readable Data

### Introduction

Since much of the social science data, especially in the United States, dealing with the 1960s and 1970s is computer based, the preservation and accessibility computer-readable data demands the immediate attention of social science researchers. If there is to be significant social science policy research in the future on the 1960s, 1970s, and 1980s the preservation and assessibility of the computer-readable data so vital to this research must receive adequate attention now.

In this context, therefore, this paper treats the problems of preservation, storage, and access to computer-readable processes-produced data from the vantage point of the National Archives and Records Service. Part one sketches an historical overview of the experience of the National Archives and Records Service in this area. Included is a discussion of the application of computer processing information technology to the creation, use, and storage of Federal records since 1951, and how the National Archives and Records Service has responded to this new technology. Part Two examines certain problems, policies, and procedures in the preservation and access of computer-readable data.

### Computer Technology

The development of computers in the United States really began during World War II when a crude electromechanical computer called MARK I was developed to generate firing tables for the U.S. Navy. Toward the end of the war the Electronic Numerical Integrator and Computer (ENIAC), which was the first electronic general purpose computer, went into operation at the University of Pennsylvania. After World War II the U.S. Army and the Manhattan Engineering District (which had the responsibility for developing the atomic bomb) supported the development of



computers. Also, the Office of Naval Research formed a computer section to study the impact of computers on science and technology. During this same period the National Bureau of Standards developed a computer called SEAC for Standard Eastern Computer. However, the first commercial development of a general purpose computer was the result primarily of the work of J. Pressley Eckert and John Mauchly who had developed ENIAC at the University of Pennsylvania. They formed a company called the Eckert-Mauchly Computer Corporation (later acquired by Sperry-Rand) and secured a contract from the Bureau of the Census to produce a general purpose computer<sup>1</sup>. This computer, which was called UNIVAC for Universal Automatic Computer was installed and used to process the census returns of 1950.

Even though by today's standards UNIVAC is quite crude, at that time its usefulness in storing and manipulating enormous amounts of information was quickly recognized. In 1952 there were five computers in the Federal government. During the next decade this number grew to 1,006. By 1977 the total number of general purpose digital computers being used by the Federal government exceeded 9,648. It is estimated that the annual cost of computer hardware alone is between six and ten billion dollars.

More important than the number of computers is the amount of information created and processed by them. Currently, Federal agencies use between 8,000,000 and 9,000,000 reels of computer tape (2400 feet long) to store and process information<sup>2</sup>. Furthermore, approximately 40 per cent of the information processed by Federal agencies is computer based and is increasing each year.

Almost every Federal agency now uses computers to generate, store, and manage information. While much of this information is routine bookkeeping of little interest to social science researchers, a significant amount does provide basic research materials. For example, between 1963 and 1972 the Department of Defense supported a number of computerized information systems that analyzed a wide variety of data regarding the United States war effort in South Vietnam. Included in these systems are data used to measure the effectiveness of the army's pacification program; to provide detailed information on terrorist activities; to identify patterns of North Vietnamese and Viet Cong activity and to serve as an intelligence base for military decisions; and to assess the ecological, physiological, economic, social, and military effects of the herbicide program<sup>3</sup>. More than thirty files comprise this unusual data

<sup>1</sup> Shelton, Bill, and Duncan, Joseph W., *Revolution in U. S. Government Statistics, 1926-1976* (draft manuscript from the Executive Office of the President, Office of Management and Budget), Ch. 4.

<sup>2</sup> Magnetic tape is the primary storage medium for computer-readable records. Punch cards and paper tape seldom are used for storage. Disc storage devices are used principally for records that require virtually on-line access.

<sup>3</sup> For more information see *Catalog of Machine-Readable Records in the National Archives of the United States* (1977).

base. Other Federal agencies that create substantial computer-based information include the Internal Revenue Service, the Department of State, the Bureau of the Census, the Bureau of Labor Statistics, the Department of Agriculture, the Department of Health, Education, and Welfare, the Environmental Protection Agency, and the Energy Research Development Administration, to name only a few.

One interesting aspect of how Federal agencies use computers is management of information systems. In the Department of State a system now in operation annually receives and stores by computer more than 500,000 cables from U. S. embassies and consulates. The messages are indexed by computer and assigned a reference number. Since they are arranged in the order they are received, easy access to a particular cable or cables on a specific subject is possible only through an on-line computerized index. Equally interesting is the computerized information system the Watergate Special Prosecution Force used in conducting its investigation of the so-called Watergate affair. This system contains abstracts of testimony before the Senate Select Committee (the Ervin Committee), the Grand Jury, and office interviews. Some seven data items such as name, date, type of incident and the like were flagged so that investigators could retrieve abstracts meeting certain criteria. For instance, all of the abstracts of an individual's testimony or all of the testimony in which a single person or a combination of people was named could be retrieved. This system could be refined even more by restricting retrieval to a certain time period or a certain event.

## The National Archives and Computer Technology

This brief sketch of the growth of the application of computer information processing technology to the creation, use, and storage of records within Federal agencies is incomplete without a review of the National Archives' activity in this area. Until the early 1960s the National Archives was doing very little. Indeed, the prevailing view was that IBM punch cards, then the primary storage medium, were non-record and redundant since tabulations derived from them existed in printed form. This accounts for a decision in July 1936 to dispose of some eight million cards which contained information from the Census of 1930.

Happily, this view underwent considerable modification in the 1960s. In 1963 the Social Science Research Council, concerned about the increasing quantity of social science data in machine-readable form, appointed a Committee on the Preservation and Use of Economic Data. This group registered its concern about the preservation of economic data stored on punch cards and magnetic tapes in Federal agencies. At about the same time several people at the National Archives became quite concerned about the status of machine-readable records in Federal agencies. One of

them, Meyer Fishbein, urged that key persons from the National Archives meet with the SSRC committee<sup>4</sup>.

One result of the discussions was an examination of the possibility of the National Archives establishing a data center to handle the growing volume of machine-readable records. This undertaking coincided with a survey of statistical data on punch cards and magnetic tapes in 13 selected Federal agencies conducted jointly by the Bureau of the Budget and the National Archives. The study, which was completed in 1964, concluded that it was impractical to consider establishment of a Federal data center. Instead, a far more urgent problem requiring immediate attention was that of bringing the disposition practices of Federal agencies regarding machine-readable records under Federal regulation. The survey of the 13 selected agencies demonstrated that current policies and practices did not ensure retention of valuable data either for government use or scholarly research. Despite this less than enthusiastic response of the National Archives to the concept of a federal data center, the Social Science Research Council approached the Bureau of the Budget immediately about establishing a Federal Data Center. This proposal soon became entwined in the issue of computers and the invasion of privacy and was dropped by the Bureau of the Budget.

However, the National Archives' concern about proper disposition practices for machine-readable records did not abate. Archivist of the United States, Robert Bahmer, established a committee to undertake a detailed study of machine-readable records in Federal agencies. The Committee's report in 1968 included a number of recommendations, one of which called for the establishment of a special organizational unit to deal with machine-readable records. In 1969 the new Archivist of the United States, James B. Rhoads, created a data archives branch or staff to complete a survey of magnetic tapes in Federal agencies, developed an inventory of all such files, and tapes containing information of possible long term value.

Since 1969 the Data Archives Staff, now called the Machine-Readable Archives Division, has accomplished much. A major inventory of machine-readable tapes has been completed and updated. Some 2,000 reels of magnetic tape containing records of long term value have been accessioned into the National Archives<sup>5</sup>. A reference service that includes copying tapes and providing special purpose extracts from multi-reel files has grown to the point that more than 1,200 tapes were copied for a fee in this fiscal year. A special tape storage vault with controlled environmental conditions was completed. General Records Schedule Number 20, which identifies categories of disposable and non-disposable records, was developed and substantially revised last year. Federal regulations regarding proper storage and maintenance of machine-readable records have been greatly strengthened. And in April of 1977 the Archivist of the United States authorized the Machine-Readable Archives Division to receive machine-readable records which are of high current interest but whose long-term value is uncertain. The vehicle for accomplishing this mission is the Cen-

<sup>4</sup> Data Archives Staff Report, August 1970, Washington/DC 1970, p. 1.

<sup>5</sup> Catalog of Machine-Readable Records in the National Archives of the United States.

ter for Machine-Readable Records. Discussions with several Federal agencies that produce machine-readable records of interest to social science researchers are now underway that could result in the Center serving as a kind of clearing house for information regarding which Federal agency has particular files. Furthermore, the Center will, where possible, seek from Federal agencies machine-readable records that are of interest to the social science research community and make them available. In the long run this could be one of the more significant activities of the Center for Machine-Readable Records.

While much still remains to be done, it is clear that the National Archives has established a program whereby a wide variety of computer-readable records of the Federal government will be available to researchers in the present and the future. Certainly this is no mean accomplishment, and it documents the commitment of the National Archives to the preservation and dissemination of computer-readable records.

## Appraisal of Computer-Readable Records

Since the National Archives lacks the resources to preserve *all* of the computer-readable records that Federal agencies produce, the process of selecting those computer tape files to be preserved is important. The experience of the National Archives for the last twenty-five years or so in dealing with textual or printed records for Federal agencies is that only about three per cent of them eventually are preserved. If this rule of thumb is applied to computer-readable records then the national Archives faces the prospect of preserving between 240,000 and 270,000 reels of computer tapes as of 1977<sup>6</sup>.

The key problem here is how to identify this three per cent or 240,000 to 270,000 reels. The Machine-Readable Archives Division has developed a records schedule that identifies categories of disposable and non-disposable computer tape files<sup>7</sup>. Through this schedule the Archivist of the United States has delegated authority to Federal agencies to destroy all computer-readable records that are classified as disposable by the record schedule. In most instances, „processing files“ which

<sup>6</sup> This percentage may be too low. Lionell Bell has suggested that more computer-readable records today will be preserved vis-a-vis preservation of comparable textual records because the computer greatly extends the power of both administrators and users to handle large volumes of data. See *The Archival Implications of Machine-Readable Records*, VII International Congress of Archives (Washington/DC, September 27–October 1, 1976).

<sup>7</sup> The official title of this schedule is „General Records Schedule 20. Machine-Readable Records“.

range from data input to update transactions, are disposable without regard to subject matter. Non-disposable records which must be offered to the National Archives are „master files“ that constitute the definitive state of a data file in a system at a given time. General Records Schedule Number 20 identifies several classes of disposable „processing files“ and non-disposable „master“ files. Generally, routine housekeeping files are „master files“ that are disposable because of the trivial information they contain. An example of a disposable housekeeping file would be an inventory of army vehicles or airplane parts. On the other hand a „master file“ containing information relevant to social science research in its broadest meaning would be preserved. Statistical and scientific „master files“ must be offered to the National Archives.

The fact that a file is not disposable under General Records Schedule Number 20 does not automatically mean the National Archives will preserve it. A number of stringent criteria have been developed that must be met before a file is accessioned. These criteria are summarized in the appraisal decision table for ADP records in Figure 1.<sup>8</sup>

The process of applying these criteria goes something like this. After the tape reaches the National Archives, we deal initially with what we call technical considerations, the first of which is the adequacy of the documentation. Federal regulations specify that tape files transferred to the National Archives must be accompanied by documentation that includes agency prepared technical memoranda relating specifically to the file, a record layout, a codebook, publications derived from the file, and two completed General Services Administration tape inventory forms. If absolutely essential documentation such as a codebook or a record layout is missing and can not be reconstructed, the appraisal terminates and the tape is returned to the agency for disposal.

Usually, the essential documentation is intact and the tape is then physically checked for readability, which is the second phase of the technical considerations. This means mounting the tape on a drive and reading it. Sometimes temporary read errors are encountered that can be eliminated by passing the tape over a tape cleaner or mounting the tape on another drive. When permanent read errors are encountered, the decision about readability depends both upon the scope and magnitude of the errors as well as the basic value of the file. For example, a few unreadable blocks do not seriously diminish the technical quality of the file. On the other hand, if more than five percent of the blocks are unreadable, then the file would be considered unreadable<sup>9</sup>. However, no hard and fast rule is applied since even an incomplete file of major substantive importance still could be very valuable.

<sup>8</sup> Figure 1 is reproduced at pp. 470–471.

<sup>9</sup> This varies, depending upon record length, block size, and the pattern of error distribution. The same error in every block would be handled differently than random errors. Also, the importance of the records themselves would be another consideration. One way to handle permanent errors is to delete the block from the file and print it out for inclusion in the introduction to documentation that is prepared.

At the same time the readability of the file is determined, we also obtain a record count and a five block dump and create a standardized 1600 bpi archival copy and a reference copy. (See step 7 in the decision table.) These two copies are placed in our vault while the archival qualities of the file are evaluated.

The major archival consideration is the legal, evidential, or informational value of records (see step 9). Few computer-readable records of Federal agencies impinge on the legal rights of citizens and the Federal government. The evidential value of records refers to documenting significant agency policy decisions and the agency's mission accomplishments. Although computer-readable records seldom provide this documentation, the manner in which data is collected and used can reveal a great deal about an agency's perception of its missions.

The concept of informational value refers to the residual value of records after agency needs have been satisfied. Or to put it another way, the value of such records is that the information they contain can be analyzed in ways and for purposes other than those for which the agency originally collected the information. A trite but cogent example is the U. S. decennial census which is conducted every ten years (every five years beginning in 1985) in order to count the population and to summarize trends about the social and economic well-being of citizens. For all practical purposes, once the Bureau of the Census publishes its multi-volume report, the information has little value for agency purposes. However, to social scientists and genealogists, and other researchers, the value of these records lies in the information they contain about particular persons, groups, situations, events, and the like.

The informational value of computer-readable records is proportional to their level of aggregation. For example, a summary of census data at the enumeration district level is far more valuable to researchers than a summary of county level census data. Correspondingly, census information at the household level is more valuable than a summary at the enumeration district level. This rule is that while you can never disaggregate summarized data (reduce grouped data to individual data) you can always aggregate individual level data to the desired summary level.

A file's potential for linkage with other data is another consideration of informational value in determining whether the National Archives will preserve it. Usually, records arranged at the lowest reporting unit (individual person or individual business firm) have considerable linkage potential. Common attributes such as place of residence, occupation, age, and sex (if they share similar codes) permit the linkage of groups with similar attributes.

An assessment of the importance of the subject matter the records cover is as important as an evaluation of a file's potential for further processing and data linkage. Subject matter importance is defined in terms both of the interests and concerns of researchers today as well as the anticipated interests and concerns of researchers fifty years from now. Obviously, this kind of evaluation requires an understanding of a wide variety of research trends. Also, it involves considerable „luck“ since the accurate prediction of future social science research trends (and by definition the data required to support this research) is at best an educated guess.

If the decisions in steps 8, 10 and 11 of the decision table are affirmative, then

attention turns to data validation. This step involves a manual comparison of the codebook and record layout specifications with the five block dump mentioned earlier. If this comparison reveals any inconsistencies, values not noted in the codebook, or missing data they are noted and included in the written appraisal report<sup>10</sup>.

Data validation also involves consideration of the reliability and validity of the data. Since records of informational value probably will be used in ways and for purposes other than those for which the agency collected the data, careful attention is paid to possible biases. This is particularly true of data collected for regulatory purposes. Frequently, data validation will reveal the existence of data imputation where estimates have been substituted for missing responses or incorrent figures. Generally, data imputation occurs when consistency checks are made during edit runs prior to creation of a master file. Unfortunately, there is no simple or inexpensive way to identify where specific data imputation has ocured, although the overall process itself can be documented. An even more complex data imputation problem that data validation may uncover is the extent to which estimates have been made in working with disparate data to accord with appropriate definitions, to fill gaps in coverage, and to obtain statistical comparability among geographical units over time. Given the way this kind of data imputation occurs in, say, the Bureau of Economic Analysis of the United States Department of Commerce in preparing estimates of personal income for local areas, there is no audit trail, so to speak, that can be followed. Only the broad outlines can be discerned as inquiries are made about how the file was created.

It is important to note that the process of data validation involves no corrections or cleaning of the data. The responsibility of an archivist is to record for potential users all of the deficiencies and limitations learned about a particular file while in the process of assessing its long term value. It is the user's responsibility to determine the extent to which the reliability and validity of the data in question meet his research goals.

Even though at this point in the decision table (step # 22) a file may be acceptable, arrangement and accessibility of the data along with estimated preservation costs must be weighed before recommending accession into the National Archives.

<sup>10</sup> Typically, this requires 10 to 15 hours for a file with few problems and includes preparing an introduction to the documentation. This introduction contains an evaluation of the technical quality of the file. Incorrect data codes and missing values are noted. If permanent read errors were encountered, then a printout of that block(s) is included along with a statement as to the probable cause. There are automated data verification programs, but the amount of time required to prepare the input data when the codebook is not machine-readable can be extensive. While manual data validation can be tedious, it does have the virtue of giving archivists a „feel“ for the data. For a useful discussion of checking for errors see Roistacher, Richard A., A General Consistency Check for Machine-Readable Data, in: Sociological Methods and Research, Vol. 4, No. 3 (February 1976), pp. 301–320, and Hammer, Michael, Error Detection in Data Base Systems, in: Proceedings of the National Computer Conference 1976, Vol. 45, pp. 795–801.

Arrangement refers to the internal data structure of the file while accessibility refers to whether or not the file is software dependent or is in some non-standard character code (such as XS-3 that the Bureau of the Census uses).

Increasingly, Federal agencies are employing data base management systems such as NIPS and System 2000 and special purpose statistical analysis packages such as SPSS and OSIRIS, to name only a few. From the viewpoint of the National Archives a file embedded in any of these systems or packages is software dependent in that it can only be processed in a computing environment that supports the system. This is not a major problem with regard to SPSS or Osiris; however, it is with regard to NIPS, which IBM no longer supports, and System 2000, which is proprietary. Since the policy of the National Archives is to preserve software independent files, this means that conversion is necessary. Our experience is that such conversion is expensive. For instance, the military files dealing with South Vietnam (discussed earlier) were embedded in a NIPS format then transferred. Since only those users who had access to a computer system that had NIPS could process the files, we concluded that the value of the files was sufficient to justify an expenditure on the order of \$ 400 per reel.

The problem with arrangement is less critical in that usually for ease of processing variable length records are formatted as fixed length records. The result is considerable padding with zeroes that expands the physical size of the file. Since storage space is at a premium, data must be compacted as much as possible in order to reduce the number of reels in storage. This compaction also can work to the advantage of users who may have to pay for only one reel of tape rather than two reels.

The cost of preserving computer-readable records with existing technology is such that it can not be ignored in determining whether or not a file will be accessioned into the National Archives. Therefore, we must exercise even greater care in the future when selecting files to be accessioned. Naturally, this will encourage a conservatism that can result in preserving only files for which high research interest is demonstrable. Even though it is too early to predict its full impact on the National Archives, I do think that in the long run it will not be possible to accession the wide variety of information implicit in the concept of process-produced data. It is possible, given this constraint of preservation costs, that fifty years from now the computer-readable holdings of the National Archives will serve only a handful of researchers whose interests coincide with the narrow scope of our holdings. In effect, therefore, the exercise of such care and caution in selecting records to be accessioned could yield a very undesirable outcome; the preservation of files that may be at best of marginal value fifty years from now and the destruction of files that fifty years from now would be considered very valuable. It is conceivable that serendipitous data bases compatible with the questions of researchers in 2027 will be scarce, given the care and caution with which the National Archives must operate in this area.



## Preservation

Presently, the National Archives preserves computer-readable records on magnetic tape even though it is not accepted as a permanent archives storage medium. Computer tape on which magnetic signals have been recorded is a highly fragile storage medium that is vulnerable to powerful magnetic fields and can deteriorate rapidly in an unstable environment<sup>11</sup>. Tapes written at 1600 bpi and stored under optimum conditions should be reliable for ten years. However, in order to achieve these optimum conditions it is necessary to eliminate the major causes of deterioration of computer tape. These are exposure to radical changes in environmental conditions, dirty tapes, and fluctuations in tape tension in the rewind mode.

The National Archives maintains a fireproof vault in which some 20,000 reels of tape can be stored at a temperature of about 68° Fahrenheit and 50 per cent humidity<sup>12</sup>. A hygrothermograph in the vault monitors continuously the temperature and humidity. A backup cool air unit is activated if the temperature in the vault becomes excessive.

Dirty tapes occur when debris is deposited during the manufacturing process or when they are passed over dirty read/write heads. This problem is dealt with in part by insisting that computer operators clean the read/write heads regularly and write only on new certified tapes that have been cleaned. The latter is particularly important since the manufacturing process tends to leave debris on the tape surface that can cause temporary read/write errors. Therefore, every tape is passed over a tape evaluator/cleaner before being used. Preservation tapes are passed over a tape cleaner again after being written on to remove any dirt deposited by the read/write heads.

Since tension in the stacked (or rewound) tapes layers is a critical factor in long-term storage, balanced tension while rewinding is important<sup>13</sup>. Unfortunately, the tape drives now available to the National Archives lack this feature. The „read backward“ mode could be used, but it would double the computer cost. A relatively inexpensive solution is at hand since the tape cleaner referred to above rewinds under controlled tension.

These optimum storage conditions are supplemented through writing an archival

<sup>11</sup> For an excellent discussion of tape vulnerability to magnetic fields see Geller, Sidney, *Erasing Myths About Magnetic Media*, in: *Datamation* (March 1976), pp. 65–70.

<sup>12</sup> A 2400 foot reel of magnetic tape tends to change total length about one foot for every 20° Fahrenheit change in temperature or every ten percent change in relative humidity. These effects are independent and can occur concurrently.

<sup>13</sup> Too much tension causes the tape to assume a permanent stretch or curvature which can damage the tape surface on the first read attempt or can cause uneven winding on the second rewind. Too little tension allows some layers to separate at some point in the stack and in some cases the layers will slip or fold into permanent creases that can not be read. Thus, it is very important that the tape drive not introduce sharp variation in tension as the tape is rewound.

master copy tape and a reference copy tape at 1600 bpi in EBCDIC. In addition, each copy is verified as error free. Both the master copy and the reference are placed in plastic cannisters and stored upright in tape racks. The reference tape copy is removed from the vault as needed to process reference requests. The master tape copy leaves the vault only for a periodic check of readability.

A three percent statistical sample of tapes is tested each year for readability. This test consists of mounting the tapes on a tape drive and reading them. It is likely that most read problems will be temporary read errors which can be corrected by passing the tape over a tape cleaner. If a tape which is part of a multi-reel file has more than five temporary errors the entire file is checked. Tapes with a significant number of temporary or one or more permanent errors are copied onto new tapes and verified as error free. Plans now call for copying every tape onto a new tape after ten years of storage. These procedures cost less than rewinding tapes every six months. Also, the use of backup tapes (master copy and reference copy) provides an added dimension of protection.

Even though the National Archives and Records Service does not recognize magnetic tape as a permanent archival storage medium research, development now underway suggests that the time is near when permanent storage medium for computer-readable records will be available. While no firm position has been taken, several criteria have been developed that could be used in assessing the capability of mass storage technology to meet the Archives' needs for permanent storage medium.

An acceptable permanent storage medium for computer-readable records in the National Archives should meet several criteria. Absolutely essential is a storage mode in which under typical conditions, the recording signal will not degrade over time. This requires a non-erasable or read only capability. To achieve this criteria the recording signal must affect an unalterable change in the storage medium. A second criterion is a non-volatile storage medium that could be treated much the same as ordinary library books. Maintenance costs, such as that associated with a controlled environment and periodic refreshing of the storage medium, would not be necessary. A third criterion is immediate verification that an error free copy has been made. The importance of DRAW (Direct Read After Write) is significant when a large volume of records is processed and it is too costly or otherwise impractical to verify an error-free copy later. While it is expected that most computer-readable records in archives storage may not be used often, it is most desirable to have a storage medium that can be read repeatedly with little or no degradation in either the signal or the medium. A fourth criterion, therefore, is that the read capability must not involve physical contact with the storage medium like that of tapes passing over read/write heads. High packing density at a low cost is the fifth criterion. Closely related to high packing density is a high data transfer rate, which is the sixth and last criterion. A data transfer rate on the order of one megabyte per second is necessary for efficient processing of voluminous files (some of the multi-reel files in the National Archives contain as many as eighteen 1600 bpi reels). The significance of a high data transfer rate becomes increasingly critical as the volume of computer-readable records grows.

Currently, magnetic recording is the principal means of storing computer-readable records. For more than twentyfive years magnetic technology has dominated the storage of digital data<sup>14</sup>. Magnetic tape offers low cost, direct read after write, erasability, and a high density, especially 6250 bpi tape. Magnetic disks, on the other hand, permit random access, along with some of the other features of magnetic tapes. Nevertheless, magnetic recording has serious deficiencies in archival storage, not withstanding a number of significant developments in the last year or so<sup>15</sup>. Low maintenance, non-erasability, and no physical contact with the medium are almost impossible to achieve magnetically. The one exception to this is magnetic bubbles which do not involve physical contact with the medium while reading or writing.

Recent developments in non-magnetic storage media research suggest that optical recording offers the greatest potential for meeting archival storage requirements for machine-readable records. While several optical recording devises and approaches are still in the development stage, the technology generally consists of a powerful light beam that can be modulated to produce a stream of bright and dark spots representing the 1's and 0's of digital data, a lens or objective to focus the beam, and a sensitive storage medium. A laser which can operate at about  $10^6$  pulses per second with the power of one watt is sufficient. Optic lenses that can focus a light beam down to a dimension of less than one micron of surface area are available. This corresponds to  $6 \times 10^8$  bits per square inch or the equivalent of one fully packed 2400 foot reel of magnetic tape written at 6250 bpi. Storage media include photographic emulsions and film with a thin reflecting coating of metal<sup>16</sup>.

<sup>14</sup> Freeman, David N., The New Mass Storage Systems, INFO 76 Conference, Chicago 1976.

<sup>15</sup> Dollar, Charles, Problems of Magnetic Recording in Archival Storage, in: Digest of Papers, Spring COMPCON 77, IEEE Annual Meeting, San Francisco 1977, pp. 28-80.

<sup>16</sup> Chen, Di, and Zook, David J., An Overview of Optical Data Storage Technology, in: Proceedings of the IEEE, Vol. 63, No. 8 (August 1975), pp. 1207-1212; Gaylord, Thomas K., Optical Memories: Filling the Storage Gap, in: Optical Spectra (June 1974), pp. 29-34; A Report on the Archiving of Binary Computer Data on Microfilm, Commissioned by the United Kingdom Central Computer Agency, in: Informational International, Inc. (February 1977); Kaczorowski, Edward M., Optical Mass Storage, Digest of Papers, Spring COMPCON 77, San Francisco 1977, pp. 33-35; Kenney, G., et al., An Experimental Optical Disc Data Record, in: Technical Note, No. 144, Philips Laboratories (July 1977).

## Dissemination

The National Archives and Records Service preserves computer-readable records in order to make them accessible to users — now and in the future. Accordingly, the policy of the National Archives and Records Service is to provide the widest practical access to computer-readable records that is consistent with statutory and regulatory constraints and available financial resources.

A unit within the Machine-Readable Archives Division has the responsibility for providing reference service that include preparation of a catalog that describes our holdings in non-technical terms. The 1977 edition of the *Catalog of Machine-Readable Records in the National Archives of the United States* lists more than 125 entries that range from a public opinion survey on population growth and the future to a metropolitan Washington area transportation study. Identification of these entries is keyed to the concept of record groups rather than subject matter. As the volume of our holdings increases we expect to construct a machine-readable subject index to each entry or file in order to facilitate user access to data relating to specific subject areas. Our current practice of describing entries probably will be adequate for several more years or until the volume of our holdings makes it unwieldy.

Two general types of reference service are provided: copies and extracts. Copy reference work includes card-to-card, card-to-tape, tape-to-tape, and tape-to-print-out, although tape-to-tape copying makes up about 90 percent of the work. A policy of providing data to as many users as practical in a format that conforms to the computer hardware requirements where the data will be processed means that a user may request a seven or nine track tape (BCD or EBCDIC), with or without internal labels, and written at a density of 556, 800, or 1600 bpi. A complete documentation package for each file is available in electrostatic copies or 16 mm microfilm. Approximately 75 percent of our reference work consists of tape copying.

Generally, extract work is done for a user when only selected portions of a multi-reel file are needed. A special purpose program is written according to user specifications and an output tape with a modified documentation package is provided to users.

The Machine-Readable Archives Division does not provide special software service. Indeed, our preference is to have tape files software independent so that many more users will have access to a file. I should hasten to add that most of our files are formatted so that they can be accessed by SPSS or OSIRIS after being prepared for data input. Sometimes we accept as reference tools special software programs written for certain files. We will make these programs available to users but assume no responsibility for maintaining them. Our limited resources simply do not permit involvement in all of the problems of supporting special purpose software.

Users must pay for the services received. A tape-to-tape copy without regard to density, character set, or labels costs \$65 per output reel plus documentation. The

Figure 1:  
Appraisal of ADP Records

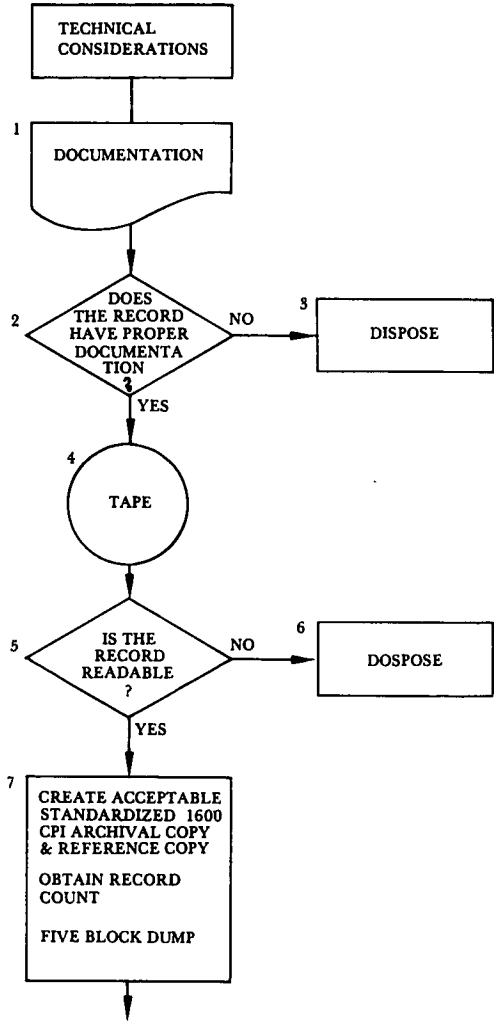
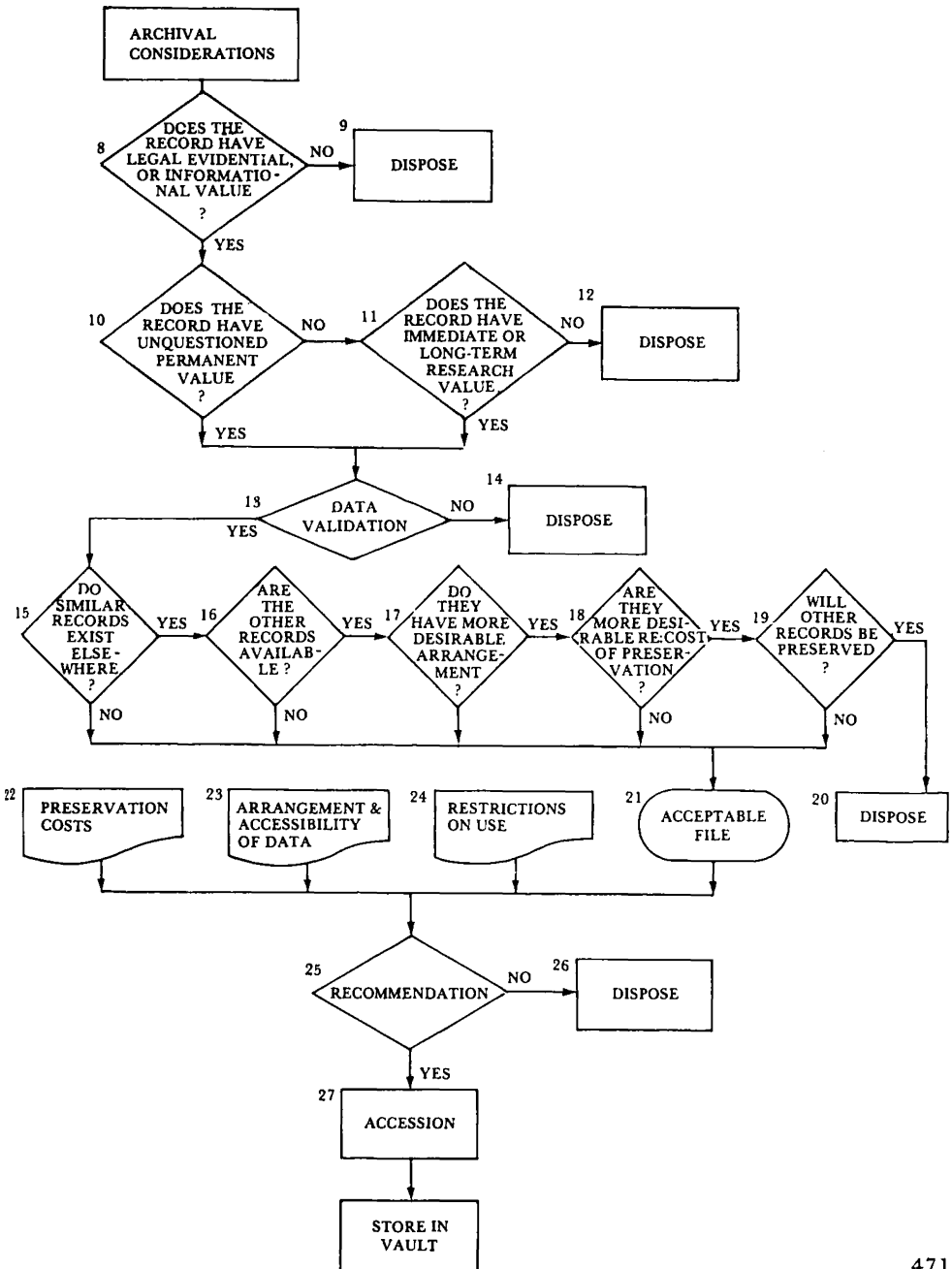


Figure 1:  
continued



fee charged for extract work is \$65 for each output reel plus \$150 per hour of computer processing time used in extracting the desired information. If the desired output is a printout the only fee charged is \$150 per hour of computer processing time. This is the lowest fee charged by any Federal agency that is required to recover costs incurred in disseminating computer-readable data.

Inevitably, when dissemination of computer networking is advanced as the future mode of disseminating computer-readable records. There is no doubt that the technology now exists for computer networking. Some 160 universities, military installations, and the like in the U.S. and Europe are now part of the ARPA Network. A number of computerized data services provide online transcontinental computing service. The primary problem from our perspective is that data transmission rates are so low that it would be prohibitively expensive to users who desire several reels of tape. Thus far, our experience is that most users of the computer-readable records in the National Archives are satisfied with the delivery of tape by the United States Mail Service. Of course, another approach would be for users to be part of a computer network which could access a subject index to computer-readable records in the National Archives. If the documentation were computer-readable and on-line with the index, a user could peruse the two, select the items of information he wanted, and then key in a request which could be processed and the output returned by U. S. mail. Our experience has been that most of our users are not sophisticated enough in this area to benefit from such access. As users become more competent in on-line information retrieval and the cost of computer networking decreases, the National Archives and Records Service will play a role in computer networking. In the meantime, we will continue to use magnetic tapes and the United States Mail to meet the data needs of our users. At the same time and as our resources permit it, we will initiate action to make it less difficult and costly for the National Archives to become a viable part of such a network.

## Conclusion

This paper has described now the National Archives and Records Service preserves and disseminates computer-readable records. A point worth repeating is that careful attention is given to the matter of identifying those files that merit preservation. The National Archives believes that considerable progress has been made in this area, especially with General Records Schedule 20. Nevertheless, both the preservation and dissemination of computer-readable records in the future will be influenced largely by technological developments. In this regard, the National Archives' experience thus far in dealing with computer-readable records is only a beginning. The experience of other national archives will provide a better picture of the magnitude of our collective problems and solutions.

## Archives and Machine-Readable Data from Public Administration in the Federal Republic of Germany

My subject is machine-readable data in German public administration. Because the experience of the federal and state archives in West Germany is very limited in this field, I will keep my remarks appropriately brief. Until now the Bundesarchiv has acquired only about fifty magnetic tapes from federal agencies and other institutions. The archives of the West German states possess an even smaller number of such tapes. This is a very unsatisfactory situation, which merits a brief explanation.

German archivists discovered computers relatively early, but their main interest for many years was limited to the use of Automated Data Processing for straightforward archival work such as the indexing of records. For example the Hessian State Archive (Hessisches Staatsarchiv) in Marburg has developed a computerized system to sort and print out the personnel data for Hessian soldiers who fought for the English government in the American War for Independence<sup>1</sup>. In the Bundesarchiv a computerized index has been assembled to maintain data files on the researchers who have been given access to our holdings and on the studies they undertake.

To reconstruct missing personnel records we have a computerized register of over 300,000 persons who are mentioned in surviving records. This file enables us to answer numerous inquiries for documentation of previous public employment or military service. Another computer project of the Bundesarchiv is documentation of the destinies of Jewish inhabitants of the Third Reich, who were living there in 1939. This latter work will hopefully be finished in about two years.

The main purpose of archival use of Automated Data Processing in the Federal Republic has been the indexing of records. Frequently the archivists have gone far beyond this task by expanding their indexes with additional evaluative information. In the Bundesarchiv, for example, a personnel data file is being assembled with information about persons who faced judicial hearing (Spruchgerichte) in the British zone after 1945. 20,000 of a total of 32,000 personnel files on hand have already been processed. Included for each defendant are, among other items, the name, birth date, religion, education, position in the National Socialist Party and its related organizations, position in public service and the decision of the court. While this is not presently available for research, it is being used for archival purposes and will eventually be made available for quantitative evaluations by historians.

<sup>1</sup> Hessische Truppen im amerikanischen Unabhängigkeitskrieg, in: Veröffentlichungen der Archivschule Marburg, Vols. 1 to 5, Marburg 1972/74/76.



The internal archival uses of ADP may partly explain, but not totally excuse, the failure of West German archivists to acquire machine-readable data produced in public administration. Actually German archivists are formulating expert guidelines for transferring German census data, on both the federal and state levels, to the archives in machine-readable form. Given our relative lack of experience on this field, numerous problems remain to be solved.

One major problem we have already confronted is the legal one. We had to ensure the right of the Bundesarchiv to acquire machine-readable data on magnetic tape. The Bundesarchiv has always had the mandate to preserve paper records as prescribed in the Registraturanweisung, the registration ordinance. In 1975 the definition of „records“, that is „Schriftgut“, was expanded to include all kinds of data irrespective of their form<sup>2</sup>. Since then we have had the clear legal responsibility for handling machine-readable data from German federal agencies.

This regulation appears at first glance to be very simple and self-evident. Nevertheless, it entails some difficulties. Above all, certain rules which are appropriate for paper records are now valid for machine-readable files as well, although such rules are inapplicable to the new technology. Paper records, for example, are generally closed for thirty years from the time the records are taken out of service. A computer file which is continuously updated would therefore remain perpetually closed. In my opinion, if we allow for the protection of individual privacy, machine-readable files could be opened for research much earlier than the current thirty year regulation permits.

The regulation which requires agencies to turn over their magnetic tapes to the archives fails to mention questions of the costs which arise in the process. The relinquishing agencies regard the deposit of their paper records as a favor. Their ever-growing mountains of paper files in storage practically force them to deposit their records in the archives. With machine-readable data we are not concerned with used-up paper, but reusable magnetic tapes. Until now the Bundesarchiv has received its few magnetic tapes at no cost, although always with the reservation that, for the long term, a solution must be found in the budget of the federal administration. A decision remains to be made in this matter.

A major problem which has been solved for the Bundesarchiv concerns the technical control and conservation of machine readable data. In January 1976 the Bundesarchiv rented a small computer system, which includes a 1600 bpi, nine track magnetic tape unit, a disk with 67.5 million bytes and four displays. Although our set-up is relatively small, it offers the most important services which we require in the handling of machine-readable data. We are able, for example, to check magnetic tapes, to copy them for conservation purposes at regularly timed intervals, and to manipulate the data fields. Individual fields can be taken out of the record. For instance, protected personal data might be rendered anonymous by eliminating the

<sup>2</sup> Registraturanweisung (Anhang I zur Gemeinsamen Geschäftsordnung I), Gemeinsames Ministerialblatt 1975, No. 27, p. 594.

name field. In order to handle very large amounts of data, especially when a high blocking factor is used, we are obliged to rely on a larger computer of the federal administration in Bonn.

A precondition for the accession of magnetic tape data is, in every case, the determination of an archivist that the data are worth preserving because they are of historical value. For non-machine-readable data, the Bundesarchiv was able to rely on guidelines set by Professors Scheuch and Schmölders in Köln. This set of rules is not applicable to machine-readable data. With this data it is rather useless to test a sample since the problem of quantity is no longer a hindrance as it was formerly in the case of traditional card files. Through the possibilities of computerized evaluations, in my opinion, data can be worth preserving, although in non-machine-readable form the same data would be regarded as expendable. In coming years it can be predicted that the state archives will take over more machine-readable data than is necessary. The state archives may justify the acquisition of tapes from computer centers in terms of providing security storage for machine-readable files. This means that state archives will also have to accept machine-readable data of limited or even no historical value in the coming years. The appraisal of these files will have to be undertaken at a later date based on experience accumulated in the interim.

Based on what I have said, one might draw the conclusion that institutions other than state archives should concern themselves with machine-readable data produced in public administration in order to prevent this data from being lost to historical scholarship. Even if I admit that the state archives in West Germany have perhaps been insufficiently concerned with this problem, I cannot therefore accept as preferable the view that data of state provenance should be given over to institutions such as universities or other centers for advanced research. Allow me to support my position with four arguments:

1. German state archives acquired their present function about 150 years ago. Since then they have been the only institution within public administration to have a broad overview of the archival needs of all areas of the administration. This tradition allows them to appreciate the full range of historically valuable data. A research institute of university center normally specializes in a particular field and prefers to concern itself only with the data in this field. A state archive with its precisely defined responsibility possesses the prerequisites and facilities for pursuing a broadly conceived program to cover all areas of public administration.
2. In many cases the introduction of data processing does not correspond to the introduction of a new administrative task. Data which had previously been handled manually are now transferred to automated processes. Before this procedural change was introduced, written data, handled in the traditional manner (e. g. on index cards), came to the state archives. It could not serve any reasonable purpose to select a different archival destination only because the data are now computer-generated.
3. It may be objected that machine-readable data is easy to copy and can be preserved without difficulty in two or more places. It must, however, be pointed out that information from public administration, independent of the storage

medium, can only be made accessible to the public under certain conditions. These conditions are fixed in the „Benutzungsordnung“ (i. e. rules governing the use of West German archives). These rules have proven their worth until now. Some changes may be necessary for the special conditions which obtain for machine-readably data, as I have discussed previously in relation to the practice of keeping records closed for a specified number of years.

4. The legal situation which governs the transmission and use of data about individual persons, as evidenced in West Germany's Federal Privacy Act (Bundesdatenschutzgesetz), is becoming increasingly restrictive. Given the technicalities of the law and the difficulties inherent in acquiring and utilizing machine-readable data bearing upon individuals the problems will multiply in the varying circumstances of institutions other than state archives.

My last few statements might be understood as suggesting that a conflict of interest exists between the state archives and the academic research facilities. This impression would be wrong. I believe that the state archives and all institutions for higher learning and research have common interests. Their common interests, however, are best served, in my opinion, by a separation in their functions. The state archives have as their primary task the acquisition and conservation of machine-readable data from public administration and only secondarily to pursue their own scientific research. Universities and research institutes can surely only profit by concentrating their efforts on data evaluation and research. Here the differentiation is very clear.

In closing I would like to mention the problem of handling the immense amounts of data of non-state provenance. State archives can by no means claim exclusive competence for acquiring and keeping this data. In this field it should be possible to arrange satisfactory forms of cooperation which will be fair to all interested parties.

## How to Teach Data Producers „The Noble Art“ of Data Documentation

### 1. Introductory Remarks

In the good old days, preservation, storage and access problems in the social science area were successfully solved by the library and the document archive:

Before the arrival of the computer, these giant data-collection agencies cooperated without great difficulty with the data-storage institutions: the tables and the analyses published by the statistical bureaus were stored in libraries and the original data sheets (census sheets, register protocols, and the like) were with some regularity transferred to the established archives<sup>1</sup>.

In the article on Data Services in Western Europe quoted above, Stein Rokkan claims that „the inertia of the traditional institutions“ in adjusting to the storage and display demands after the computer revolution created an unsatisfied demand for mass data in computerized form; the Data Services tried to bridge the gap between production and distribution vis-a-vis the social science community.

In this paper, we shall touch on preservation, storage and access problems from the point of view of such data service organizations: What are the main obstacles to secondary analyses of the vast and ever-increasing holdings of machine-readable data, and which remedies can secure a better utilization in the coming decade.

<sup>1</sup> Rokkan, Stein, Data Services in Western Europe. Reflection on Variations in the Conditions of Academic Institution-Building, in: American Behavioral Scientist, Vol. 19, No. 4 (1976), p. 445. This issue of the ABS deals with the „data archive movement“. The academic data archives, data services, data libraries etc. will in the following be referred to by the term „data service organization“ or just „data organization“. In Europe, 7 data organizations with national coverage cooperate in CESSDA (Committee of European Social Science Data Archives); internationally, a dozen data service organizations have just established IFDO (International Federation of Data Organizations).

## 2. What is Documentation

In this section, we shall postulate that a minimum requirement for closing the gap between the data producer and the secondary analyst is a high standard of data documentation. The documentation items are listed after the following outline of a subset of the obstacles to secondary analysis<sup>2</sup>.

The *data producer* (in relation to the secondary user: *donor*) can be reluctant or even unwilling to place data at the disposal of secondary users. The donors may claim that analysis by „outsiders“ is restricted to protect the individuals registered in a file, or that s/he is afraid that secondary analysts will misinterpret the data due to a lack of background information. The former argument is real (consider the Data Law issue) for files with information on individuals, but there are measures (aggregation, anonymization, scrambling) that soften the argument; the latter concern that the secondary user may misinterpret the data is hypocritical, making a virtue of a sin of omission: if the data were aptly documented the risk of misinterpretation would be negligible. Sometimes one has the feeling that a few unspoken considerations underlie the reluctance of data producers to disseminate data: fear of a critique of methodology or even challenges of reported findings; intentions to maintain an information monopoly; and so on. We shall return to these considerations below.

The *data user* (in relation to the data producer: *secondary analyst*) has a scientific problem area that s/he wants to investigate by means of quantitative methods. The search for relevant data is difficult due to lack of information about existing data holdings. Research libraries in Europe do not catalogue machine-readable data holdings, and neither public (e. g. statistical bureaus) nor private (e. g. individual researchers, market research organizations) data producers have been eager to catalogue their data-holdings. In addition, professional prestige seems to be higher if the social science researcher collects new data for a specific purpose rather than using data already collected. Secondary analysis research designs are complicated, and few researchers have the combined skills in data processing (computer use), quantitative methods (statistics), and one or even several substantive fields (interdisciplinary research) required to engage her/himself in such projects. In addition, the value for secondary analysis of a dataset is sometimes reduced with the „age“ of the data.

The *data service organizations* (the mediator between data producers and data users) have collected and stored data with a considerable investment in the „processing“ of each acquired dataset. Essentially, what the data organizations do is what the data producers ought to do: „Produce“ complete *documentation for users*.

<sup>2</sup> For a more extensive discussion of obstacles to secondary analysis, see Hyman, Herbert H., *Secondary Analysis of Sample Surveys: Principles, Procedures, and Potentialities*, New York 1972. Chapter 1 deals with the issue.

This brings us back to the question: What is data documentation? In an International Association for Social Science Information Service and Technology (IASSIST) meeting in Copenhagen recently<sup>3</sup>, the working group indicated that the following list of documents/facilities were desirable from the user point of view:

### *2.1. Library Information*

Information on a data-file should be produced in a card catalogue form to be entered into the library system. Till this day, the inertia of the library systems in Europe has been so strong that data files are not considered for reference. In the U. S. the communication between data organization personnel and traditional librarians is better: Coordinated by *Sue A. Dodd* and following recommendations from the ALA Catalog Code Revision Committee's Subcommittee on Rules for Cataloging Machine-Readable Data Files, the Classification Action Group of IASSIST in North America has set standards that should be adopted also in Europe<sup>4</sup>.

### *2.2 Archive Information/Study Abstract*

Each data service organization has its own way of presenting its data holdings (Inventory). It would be very useful if other data holders (e. g. statistical bureaus, machine-readable divisions of traditional archives, research institutions) would publish guides/inventories to their holdings. Abstracting of data file contents (cp. the immense resources invested in bibliographic abstracting these years) is hardly seen in Europe outside the realm of the data organizations.

<sup>3</sup> IASSIST (International Association for Social Science Information Service and Technology) is an international membership organization for data organization and information center personnel, quantitatively oriented researchers, etc.

The user documentation items listed were defined at a workshop in Copenhagen, June 26–29, 1977. The Report appeared in the IASSIST Newsletter, Vol. 1, No. 4 (Fall 1977), pp. 7–10.

<sup>4</sup> For further details, see Working Manual for Cataloging Machine-Readable Data Files, compiled by Sue A. Dodd, Data Librarian, Social Science Data Library, Institute for Research in Social Science, Univ. of North Carolina, Chapel Hill – or write to Sue Dodd to have the IASSIST Classification Action Group working materials.

### 2.3. Study Description

The study description is the locus for all information on a data file necessary to secure correct interpretation of analysis results, i. e. a complete description of the research project or administrative process in which the data was originally collected and processed; this is probably the area where the sins of omission on the part of the data producer are most outstanding: Not only is it hard to find a full description when a data organization acquire a dataset, but, alas, we still see reports on quantitative research results containing incomplete technical report sections. Under the auspices of the former Standing Committee for Social Science Data of the ISSC a standard study description scheme was developed in 1974<sup>5</sup>, and this scheme is presently being tested further by 6 data organizations<sup>6</sup>. The standard study description scheme has been designed in a flexible way to allow for a multi-purpose use: Besides containing all background information necessary for a reliable secondary analysis of the file described, the information can be automatically subsetted for abstracting purposes; furthermore, the sum of study descriptions in a data organization can be used for mapping and methodological research as well as for information retrieval and intra-archival logging purposes; finally, the study description is used for inter-institutional exchange of information on data holdings<sup>7</sup>. It should be added, however, that the standard study description in its present version is geared to survey files, primarily.

### 2.4. List of Variables

Each data organization should produce a list of variables for every file in their holdings; this list of variables should be printed as well as machine-readable. Like the data abstracts may be automatically generated from the study description, this list of variables may be automatically produced from a machine-readable codebook.

<sup>5</sup> See Report on Standardization of Study Description Schemes and Classification of Indicators and Study Description Guide & Scheme, both edited by Per Nielsen (Sept. 1974 and April 1975, respectively). Both available from the Danish Data Archives.

<sup>6</sup> Since 1974, the ZA, Cologne, and the DDA have been testing the standard study description scheme, and Steinmetzarchief, Amsterdam, has used an earlier version thereof. Now, also BASS, Louvain-la-Neuve; Leisure Studies Data Bank, Waterloo; and ICPSR, Ann Arbor, have agreed to test the instrument.

<sup>7</sup> Inter institutional exchange of study description is now being tested.

## 2.5. Codebook

At the recent IASSIST meeting, it was indicated that the ideal codebook should contain 15 major items (the first three items referring to file level, the last 12 to be applied variable by variable where applicable): (1) title of study (file and subfile names); (2) format of the data file; (3) comments at file level, e. g. concerning application of missing data codes, special weighting features, special precautions for use of file; (4) variable identification (number, label, short name, mnemonics); (5) variable source reference; (6) variable location and length; (7) variable type (alphabetic, alphanumeric, numeric, symbolic); (8) number of decimal places (scale of measurement); (9) source statements/texts/questions/scale description/introductory statements related to responses; (10) answer code values; (11) answer code descriptions; (12) comments originating in field work and coding experience: interviewer instructions and coding instructions; (13) variable contingencies (filter, skip, control); (14) variable consistency (i. e. results of checking of variable contingencies); (15) reference to derived variables.

## 2.6. Classification/Index

Within the area of classification and indexing (on file and/or variable level) a number of different schemes are available, describing different dimensions. However, the testing of these schemes has not yet reached a level where recommendations can be set forth. For a discussion of variable level retrieval, see the description by Ekkehard Mochman<sup>8</sup> who is coordinating the European classification arena within IASSIST.

## 2.7. The Data Matrix

Even if the data itself can only indirectly be called documentation, it documents the results published by the primary investigator; the data matrix is the object that is described in the data documentation. It should be borne in mind that machine-readable data (unlike paper-carried information) can be read only of a correct de-

<sup>8</sup> Mochmann, Ekkehard, Information Access at the Data Item Level: Approaches to Indicator Retrieval from Survey Archive Data Bases, SIGSOC Bulletin, Vols.6,2&3 (1974-75). Edited by Alice Robbin, this issue of the Bulletin dealt with „The Data Library: Systematic, Structural and Process Problems of Data Access“.



scription of the data carrier and the physical and logical characteristics of representation of data on the data carrier is supplied. Single punch data (if possible in card-image format) is preferable for archiving purposes.

### *2.8. Special Publications*

Many data organizations produce publications concerning specific files. Items 2.3. to 2.5. above often make up several hundred pages for one survey. Such data documentation publications should, of course, be entered into the library system. Very often, all data documentation is machine-readable.

Even if the list of user requirements for proper documentation of data files may be biased in the direction of survey data, we shall claim that it is readily generalized to process-produced data files (administrative records). Therefore, per definition, the production of written (and preferably machine-readable) descriptions following the eight-item outline above<sup>9</sup> is data documentation. Thus we can proceed to the „why“.

## 3. Why Data Documentation

In the preceding section, we postulated that data documentation is an indispensable prerequisite for secondary analyses. In this section, we extend the argument further by outlining some considerations which support extensive data documentation even in the primary analysis phase/administrative process.

### *3.1. Reliability Aspects*

The late Sir Cyril Burt, an outstanding English psychologist for decades, seems to have involved himself in research reporting of a dubious nature (hostile people call it research swindle): In his old age he has reported statistical findings based on data that (presumably) have never been collected. In *New Scientist* a survey among

<sup>9</sup> The whole data documentation standardization issue is now being addressed also within IFDO.

readers soliciting known cases of „scientific“ reports based on erroneous data resulted in 199 acceptable responses. Apparently, such examples of „research“ on the continuum from swindle to more or less conscious but severe misinterpretations in quantitative social science research are legion, and in a number of cases political directions have been based on the research; and probably the known cases constitute only the tip of the iceberg<sup>10</sup>.

Vis-a-vis the public, Danish newspapers and Radio/TV persist in their interpretation and speculations based on statistically insignificant ups and downs of political parties reported from opinion polls. In many countries, statistical figures reported seem to be dependent on power elite preferences rather than enumeration procedures.

Social scientists are responsible, and probably the hoped for impact of ethic codes is overoptimistic. In the „hard“ sciences, research results can frequently be verified by repetition of the experiment; in the „soft-data“ social sciences, measuring unique (i. e. non-repeatable) phenomena, control of results and conclusions are possible only if data and full documentation are readily available to all.

In most research projects, certain „make-up“ processes are necessary to produce a decent report from less decent data; in this paper, we are too polite to discuss the issue in further detail. However, there are cases where a random number generator would be the cheapest and most harmless data source; the quality of the research is not determined only by the sophistication of the analysis programs applied.

### *3.2. Methodological Aspects*

In effect, the data documentation requirements outlined in subsections 2.3. and 2.5. above are trivial check-lists only. All this information must be in the head of the primary investigator, the only problem being to get it down on paper (better: into the computer) in a structured way. The structuring of this methodological information may have a positive effect on research quality — as may the consciousness that secondary analysts may criticize. The reported failures and errors may have an unexpected cumulative effect on social science research quality over time.

It should be stressed that we are not statistical purists: We do not advocate that, for example, data cleaning be continued until the last invalid code has been corrected; we do advocate, however, that all known imperfections be reported, that warnings be provided of all lacunae. In short, we ask that all methodological considerations and decisions be reported at the time and place they are relevant. Computer memory is more persistent than human memory.

<sup>10</sup> For a fuller description of these events, see Science 26 (1976), resp. New Scientist, September 2 and November 25 (1976).

### 3.3. *Economic Aspects*

In the area of sample *surveys*, the Technical Section of the Danish National Institute of Social Research has found it cheaper to clean and document data files for general use before the primary analysis is started. The time and money savings in cases where reports on new issues can be based on existing well-documented files is considerable. The potential accumulation of new knowledge based on several well-documented files (longitudinally, cross-sectionally, cross-nationally) is, in the end, an economic benefit. This is an area that data service organizations are now moving into.

In the fields covered by *statistical bureaus* the production of printed statistical information is still considered the main function. Despite the fact that the statistical publications are produced by means of a computer, we know of several examples where, for example, researchers have punched the figures from the printed publications rather than requesting the machine-readable file. The level of servicing of machine-readable files from statistical bureaus to external users is relatively low.

### 3.4. *Historical Aspects*

The machine-readable divisions of the traditional archives acquire an increasing amount of tape reels containing process-produced and statistical information. With the automated text-processing revolution of the coming decade these information repositories will find themselves busy. Given the present documentation standards of public registers, the problems (from an archival point of view) in handling on-line real-time databases, and the scarce allocation of resources to the storage-display functions in National Archives, there is a risk that future historians will find problems in dealing with data from the seventies.

## 4. Who Produces the Documentation

Let us make explicit what has been implicit above: It is the data collector who produces the data documentation, not the storage-display agent (archive, service-section of a statistical bureau, or data service organization). The latter may develop standards and set directions, but the former has the information necessary for good data documentation.

Long articles could be written on the many cases in which it has proven impossible to re-use existing data. Storage of undocumented machine-readable data is organized waste of time and money.

## 5. Summary on the HOW

Having touched on the what, why and who, we shall now try to sum up how new attitudes and behavior on the part of the data producer may be implemented. The list of directions below by no means claims to be complete in any sense; rather, the directions listed are the ones that are envisioned or being implemented within data organizations.

### *5.1. Focussing on the Issue*

A necessary condition for better documentation standards is to focus much more sharply on the issue than we usually do: Within social science periodicals, such „technical“ discussions are rare; in the textbook area, we have many monographs on sociological method where the technicalities of computer-data-handling are only mentioned, we have other textbooks on advanced analysis techniques — but we have few books on computer usage for data entry, editing, cleaning, and documentation. Within the area of software the situation is similarly biased: There are many programs (packages) for statistical analyses, structure searching and even content analysis, but only a few programs for trivial data processing tasks — and generally poor facilities for integration of documentation with the data. (ZAR and OSIRIS are exceptions from this rule as both work on text-data; typically enough, these program systems have been developed within the biggest data organizations )<sup>11</sup>.

Although it seems necessary that „data pushers“ in the data organizations take the lead in this process of change of attitudes, some help is needed from the social science professionals. (In Europe, the two groups happen to overlap which is an advantage). Within the university world, student reports, doctoral theses and other reports on quantitative research the quality of methods applied in the data handling process should be considered in awarding merit. The collection of a solid data base accompanied by published data documentation is (even if the associated analysis report may be poor) a contribution to social science resources; a research report based on data analysis is nothing without proper technical reporting because interpretation of analysis results is impossible.

<sup>11</sup> The OSIRIS program package for data handling and statistical analyses is the only one among the larger analysis packages used in Europe that has unlimited and flexible built-in facilities for an integrated data documentation; the extended OSIRIS dictionary-codebook is input to the ZAR retrieval system used for text-retrieval by the ZA and DDA. (The two systems were developed by ICPSR/ISR, Ann Arbor, resp. the ZA, Cologne — in the later phases in cooperation with the DDA).

### 5.2. *Giving up Data Monopoly*

There is still the monopolistic attitude to overcome — and this attitude is seen in private organizations, among individual researchers, and in public research and statistical bureau alike. In the United States, data can be bought whereas in Europe the transfer of data is based on good-will. Even in Eastern Europe, the private-property-attitude concerning data prevails.

Explaining free data exchange as a quid-pro-quo arrangement from which everybody wins and nobody loses; providing that the anonymity of the respondents will be guaranteed; giving the data producer control of access — all of these good arguments do not always convince the data producer that he should deposit data with a data organization. Again, we end up with the data documentation obstacle: Either the data producer is afraid that the data organization will take too much of his/her time when processing the data file — or the data producer simply does not want to show dirty underwear in public; in the latter case it is a poor comfort that this grey shade is the rule rather than the exception.

### 5.3. *Offering Services*

From the point of view of data organizations, the most efficient acquisition policy may be to offer better services in terms of textbooks, software, technical aid, methodologically oriented courses, etc. Seen from the economic angle, it is probably less expensive to offer such services to primary investigators than to do a lot of documentation and processing of the data after having acquired the file from the donor.

### 5.4. *Enforcing Free Data Flow*

Science foundations and other research granting agents may make it a condition for grants that quantitative data be made available, e. g. by being deposited with a data organization. The Danish SSRC does this, and the clause applied adds that the data should be properly documented according to standards set by the Danish Data Archives (which itself is an SSRC initiative).

A different initiative has been adopted by the *Journal of Personality and Social Psychology*: The authors of articles shall at the latest five years after the publication give access to underlying data files.

### 5.5. Education

No doubt only the next generation of social scientists will learn fully how to use the computer. Consequently, it is very important that social science students be offered methodological courses where computer-use, statistics, and application of the theories of their discipline are integrated. The development of instructional data packages some of which can be used at the undergraduate level is an important contribution in this area<sup>12</sup>.

### 5.6. Infrastructural Considerations

For future social scientists exploitation of the vast data bases of statistical and administrative nature will be of great interest. Data Laws may be a new obstacle to secondary analysis; in some countries, government commission reports reveal the fact that the commission has had a double target: To protect individual integrity and to secure an information monopoly for the public bureaucracy. We may be building a barrier that makes infrastructural developments in the sense of a more smooth flow between public institutions and the research community almost impossible, thus adding to the historically determined inertia of giant data collection agents in the public sector.

However, Data Laws may have a positive effect on the quality of public data as well; with strict rules for error corrections and data organization and management, the documentation standards of process-produced data may well improve considerably.

After Data Laws have been passed the communication and flow of data from public agents to the research community may improve; at present, data producers in the public sector are reluctant to engage themselves in regular data transfer arrangements partly because they do not want to contribute to even stronger data legislation.

<sup>12</sup> Development of instructional data packages was started with the SETUPS series being a cooperative effort of the ICPSR and the American Political Science Association. Now, the International Social Science Council is sponsoring the development and testing also European teaching packages.



## IX. New Data Bases





## Development of Cross National Comparative Data from Social Bookkeeping Sources

A wealth of numbers is published every year by governmental, inter-governmental and private agencies. These numbers are designed to measure one or another of the social products or forces of social life. They represent the society's attempt to keep book on itself, to monitor its own internal workings. Virtually all governments in the modern world keep track of population movements, health and educational levels, investment, production and consumption, employment and prices, public expenditure and receipts, and a myriad of other areas of life. In most countries, these are supplemented by data collected by private and semi-private organizations. Some of these data come from censuses and public opinion polls and relate to individuals. Other data refer to somewhat larger aggregates such as firms or households. Still others are statements about the whole society; population and gross national product would be examples. Finally, some tell something of the distribution of a value or a good within the society; coefficients of income inequality and geographical distribution of hospital beds are examples of this. This public social bookkeeping is, of course, available to social scientists for their research and a great deal of use has been and is being made of these data.

Some caution is required in using these data, however. Data which are in fact collected and made available are no simple random sample of all data that are collectable. They certainly are not a comprehensive set of all conceivable measures of social life. They have been selected with some regard to some notion of what is and what is not relevant to something. In other words, there has been at least some implicit, partially developed theory that included some conception of the usefulness of the particular set of measures. Perhaps an example from early social statistics will illustrate this. In the United Kingdom, the work of the Royal Statistical Society in the 1830's and 1840's demonstrates some very specific goals and a clear conception of what was relevant to its interests. Data collected by these statistical pioneers had to do in great part with criminal incidence and educational levels and much of the analysis attempted to show that crime and ignorance were correlated. The obvious and frequently stated policy implication was that more attention should be given to public education.

Conflicting theories can give rise to incomparabilities in data. One example can be found in national account statistics. In determining the value of the economic product for the entire national economy, most countries of the world factor in the value produced by the service sector along with that of industry and of agriculture.

Communist ideology, on the other hand, holds that the service sector does not create new value, for these countries one has material product rather than national product and conversion to the latter requires estimates of the missing value of the service sector. Published data may also be designed to overcome differences of theoretical orientation. A nation's concept of what education is supposed to be is reflected in its school system and school attendance figures are, in turn, an indication of this concept. The number of years devoted to primary education vary widely from country to country. The types of secondary education vary even more widely. The UNESCO Statistical Office, however, has created a standardized system of reporting so that the national diversity can be reported in comparative terms. Educational data are reported for first, second and third levels each of which have been given specific definitions and boundaries. This standardization is itself based upon theoretical considerations and results when using UNESCO series are partly determined by these considerations. Even the least well thought out series is based upon some implied set of standard units of collection and makes assumptions about how these units are alike. Data, for example, may be collected for countries so that China and Luxembourg are considered equivalent cases but Liechtenstein is excluded on the grounds that it is unlike the other two. Again this is a decision that affects results of analyses done with the data.

Government statisticians and other officials are not always interested in quite the same problems as scholars. Even if they were, however, the problem would remain. The selection of things to measure and the choice of procedures to do so are dependent upon the approach one has to the problem. The decision to take entries in a column of data and make them „values“ of a „variable“, therefore, is itself a theoretical exercise. It is an interpretation. The matching of numbers to phenomena which is measurement requires careful thought. There is no substitute; no tricks or formula will suffice. The varied purposes that lie behind the original collections of the data — even if these ideas have atrophied while inertia carries on the series — may be contradictory or tangential to the uses to which a scholar wishes to put them. It is often necessary to select carefully among several available alternatives, none of which is exactly suitable to the current theoretical purposes. Sometimes it is possible to adjust available data so that they are more suitable to the purposes of the research. In every case, careful attention must be given to the peculiarities of definition and special methods of collection and examination must be made of possible consequences for the analysis.

The individual scholar then is at the mercy of what is available, although that is of considerable amount and variety. The scholarly community, on the other hand, is able to interact with the official statisticians on what is and how it is collected. In this paper, we look at how some of this interaction has taken place in economic and social data and investigate some possibilities for interaction in political data.

## 1. National Account Statistics

National income accounting was first developed in the 1930's for use with the macro-economic model of John Maynard Keynes. It specified the relationships among various components of supply and demand and made possible estimations of change in these relationships under particular conditions. For example, if in the model, tax revenues or government investment were allowed to increase, what effect would this have on unemployment or the balance of payments. The data and data format grew out of the theoretical needs of the Keynesian economists. Once begun, of course, it in turn influenced the further research to be done. The better known illustrations are the negative ones. In less developed countries, the non-market subsistence sector of the economy was simply excluded from the rows and columns of the accounts. Gross national product is a statement for the monetary sector only. Development economists came to think of the non-monetary sector as a residual, potentially productive force from which laborers could be drawn. For certain problems of industrialization, this approach seemed appropriate, at least before the large numbers of unemployed began to congregate in the urban centers of the underdeveloped world<sup>1</sup>. The point to be made here, however, is that the research approach was greatly influenced by the data available. Similarly, the productive capacity of housewives whose value was ignored in traditional national accounts has generally been overlooked in economic research. So has the subtraction from production represented by pollution. These were not in the data system so they tended to be omitted from analysis.

The system of national accounts has received virtually world wide acceptance at least in theory. To be sure, the Communist countries follow their own particular version and the data provided the United Nations by many countries includes a number of guesses among the components of aggregation. Even so, the Statistical Office of the United Nations has now spent a good many years defining and integrating a system which provides common standards that can produce comparative data for all of the states of the world. Published in *A System of National Accounts*, the system represents the main stocks and flows that occur within a national economy and between that economy and the rest of the world<sup>2</sup>. These stocks and flows are concerned with domestic production, consumption, accumulation and capital transactions and the balance of payments with the outside world. These accounts along with revaluations, an opening and a closing statement form a closed system. They also represent the main stay of international comparative economic data.

<sup>1</sup> Bienefeld, Manfred, and Godfrey, Martin, Measuring Unemployment and the Informal Sector: Some Conceptual and Statistical Problems, in: IDS Bulletin: The Neutrality of Numbers?, (October 1975), pp. 4-10.

<sup>2</sup> United Nations, Statistical Office (ed.), *A System of National Accounts*, in: Studies in Methods, Series F, 2 (1968).

New questions call for new data, however. Inequality of income and welfare has become increasingly a topic of academic and public discussion, for example. Most national accounting is reported in national or sectoral aggregates and tells little about the distribution of production, consumption and accumulation. With time, we can expect the data to begin to reflect the new interest. The Statistical Office has already begun consideration of a system of statistics on the distribution of income, consumption and accumulation that would be consistent with and complementary to the system of national accounts. The International Bank for Reconstruction and Development has recently published a compilation of available data on the size distribution of income in about 80 countries<sup>3</sup>. Presumably, if the political opposition is not too great, we shall soon begin to have data on inequalities of distribution by region, ethnic groups and social class.

Efforts are also being made to revise the system of national accounts. Perhaps one of the better known systems for extensive revision is that of the Economic Council of Japan. This system is consumption rather than production oriented. It measures the total economic production that is available for consumption. By omitting investments and deducting losses due to pollution but by including the estimated value of domestic work and of leisure time, it is possible to arrive at an index of net national welfare<sup>4</sup>.

Of the social scientists, economists have probably gone furthest in drawing up theoretical systems for analysis and data collection. These systems can be very powerful tools of analysis. They also make it difficult for scholars to break out of the mold and seek to follow newer approaches. Sociologists and political scientists have tended to be attracted by the possibilities of data systems, nevertheless, and have begun the search for systems that would aid in their analyses. The social indicators movement is the more advanced and better organized of these efforts.

## 2. The Social Indicators Movement

The effort among social scientists and civil servants to develop a systematic social reporting and analysis scheme is now about 15 years old. Its origin lies with several proposals written mostly by Americans in the early 1960's<sup>5</sup>. These ideas began to influence thinking especially in the larger Western developed countries so that with-

<sup>3</sup> Jain, Shail, *Size Distribution of Income: A Compilation of Data* (International Bank for Reconstruction and Development), Washington 1975.

<sup>4</sup> Economic Council of Japan (ed.), *Measuring Net National Welfare of Japan*, Tokyo 1973.

<sup>5</sup> Gross, Bertram, *The State of the Nation*, in: Bauer, Raymond A. (ed.), *Social Indicators*, Cambridge/Mass. 1966, pp. 154-271.

in a decade social reports, proto-reports and compendia of social statistics were being published in the United States, Germany, Britain, France and elsewhere<sup>6</sup>. The goals of these works included the bringing together and reorganizing of a number of heterogeneous and not easily accessible social and economic data that could be used for the analysis of social performance, quality of life, social changes, levels of living, and general welfare. Thus far at least, the response of statistical offices to some of the suggestions for systems of social indicators has been less than many scholars would have wished. The published series remain for the most part collections of series rather than integrated systems of indicators.

The goal of the social indicators movement is somewhat grander. It is to take social statistics both of the types already collected and of more original kinds and to integrate them into an ordering that allows comprehensive and precise statements on the state of welfare in society. Welfare is taken to be a multi-dimensional conception that goes well beyond the single dimension of productivity (as represented in gross national product) or even of consumption as traditionally measured. Sometimes it is approached as total consumption including the use of goods and services provided outside the market. Sometimes it is thought of as progress toward the achievement of stated national goals. Sometimes it is conceived as a subjective phenomenon when people are asked „satisfaction with life“ questions in public opinion surveys. A great deal of effort has gone into defining welfare or well-being since upon this theoretical consideration hangs most of the rest of the work<sup>7</sup>.

The effort to create an integrated frame of reference for the evaluation of welfare has not been based upon a single comprehensive social theory. Thus far in the development of social indicators, no Keynesian style social model has emerged. Perhaps it would be premature to try such an undertaking; it could provide too rigid a direction in which social data would be influenced to move. Completeness of coverage in a theoretically closed system is certainly not synonymous with completeness of coverage of all components of social life. At least for the immediate future, it is probably best to concentrate upon systematic collections of carefully selected series that are relevant to public policy.

A number of measurement schemes have gotten underway and even more have been proposed. We have space here to describe only a few of them. The scheme most similar to the National Account Statistics is that proposed by Sir Richard Stone and others and included in the Proposed System of Social and Demographic Statistics<sup>8</sup>. It suggests accounting for stocks and flows in the composition and growth of a population in a way very like the accounting already done for production, consumption, accumulation and balance of payments. Individuals, rather than

<sup>6</sup> Johnston, Robert, Review of New Compendia of Social Statistics and Social Indicators in Five Western Countries, in: Social Science Information, 15 (1976), pp. 349–370.

<sup>7</sup> Zapf, Wolfgang, Systems of Social Indicators: Current Approaches and Problems, in: Social Science Information, 27 (1975), pp. 479–498.

<sup>8</sup> United Nations, Statistical Office (ed.), Towards a System of Social and Demographic Statistics, in: Studies in Methods, Series F, 18 (1975).

dollars or marks, become the units of measure as they are classified not simply by age and sex but also by education, employment, health and other variables. For example, a matrix can be formed to show the change in educational levels of a population between two points in time. Columns can be determined by such categories as pre-school, primary, secondary, university, employed and retired and the population can be divided among them as of the first date. The rows, given the same labels, can represent the population as of the second date. The marginals of this matrix will give the stocks of population at each educational level for each of the two points in time. The cells will show the various flows from one level to another. This and similar matrices can be linked to matrices of government expenditure and of the national accounts to form one overarching system<sup>9</sup>.

This scheme has seemed rather ambitious to many official statisticians, especially among those in developing countries. It seems unlikely that resources will be devoted to so extensive a systematic collection in very many places. The United Nations have produced several papers designed to promote the improvement of social statistics in developing countries. Suggestions were made on how to produce more relevant and co-ordinated data and a framework for integrating social and demographic statistics was proposed based upon a selection and adaption of material from the System of Social and Demographic Statistics<sup>10</sup>.

Scholars have also used a number of less ambitious schemes. One of these is the battery of social indicators in which a number of conceptually simple measures are taken together for analysis. In concert they are thought to be more reliable than any one of them alone. Factor analysis or some other technique is sometimes used to reduce the wealth of data to manageable numbers. Index construction is employed by the United Nations Research Institute for Social Development in measuring welfare. Its level of living indexes are summary statements of several social indicators<sup>11</sup>.

<sup>9</sup> Stone, Richard, Major Accounting Problems for a World Model, in: Deutsch, Karl W., et al. (eds.), Problems of World Modelling: Political and Social Implications, Cambridge/Mass. 1977, pp. 57-81.

<sup>10</sup> United Nations, Economic and Social Council, Statistical Commission (ed.), Promoting the Improvement of Social Statistics in Developing Countries, E/CN. 3/482, New York 1976.

United Nations, Economic and Social Council, Statistical Commission (ed.), Framework for the Integration of Social and Demographic Statistics in Developing Countries, E/CN. 3/490, New York 1976.

United Nations, Economic and Social Council, Statistical Commission (ed.), The Feasibility of Welfare-Oriented Measures to Complement the National Accounts and Balances, E/CN.3/477, New York 1976.

United Nations, Economic and Social Council, Statistical Commission (ed.), Draft Guidelines on Social Indicators, E/CN. 3/488, New York 1976.

United Nations, Economic and Social Council, Statistical Commission, Strategy for Further Work on a System of Social and Demographic Statistics, E/CN. 3/489, New York 1976.

<sup>11</sup> Drewnovsky, Jan, Studies in the Measurement of Levels of Living and Welfare, United Nations Research Institute for Social Development Report No. 10.3, Geneva 1970.

Other systems of indicators are goal oriented. The Organization for Economic Growth and Development, for example, has drawn up a list of generally agreed upon social concerns as the basis of growth in quality of life<sup>12</sup>. Erik Allardt has put together a frame of reference for selecting social indicators to measure social welfare. He believes the values underlying welfare can be captured by the terms „having“, „loving“ and „being“<sup>13</sup>. Similarly the World Indicators Programme of the International Peace Research Institute, University of Oslo, has spent a good deal of effort in creating a conceptual system of indicators that would be more radical in content than other systems<sup>14</sup>.

Sociologists, however, have not created a single system of social indicators that has the extensive agreement underlying the system of national accounts put together by economists. A great deal more work must be done before there will be something like a closed set of concepts and theoretical specification appropriate to a single system. Political scientists are even less advanced along this road.

### 3. Cross National Quantitative Political Data

International compendia of official statistics have generally avoided political structures, processes and events. The System of Social and Demographic Statistics published by the United Nations avoids the subject by stating that it is too complex to be easily laid out in terms of quantitative analysis. Undoubtedly this is true, but the difficulty is in part due to a relative paucity of theoretical work by political scientists. Other reasons can be given for the avoidance of political data, however. Of the several kinds in which social scientists are interested, these are the most sensitive data. Some governments are not enthusiastic about revealing all the line items of their budgets or specifying exactly how and when some personnel changes take place. Elites in most countries are not greatly interested in publishing accounts of the inequalities of power. At the cross national level at least, ideological differences influence the decisions on which data not to collect. It is relatively easy to determine the proper goals for the economy. With a very few exceptions leaders in the modern world have attempted to obtain economic growth and higher levels of welfare. Although precise meanings of what this entails have differed from country to coun-

<sup>12</sup> Organization for Economic Co-operation and Development, List of Social Concerns Common to Most OECD Countries, Paris 1973.

<sup>13</sup> Allardt, Erik, A Welfare Model for Selecting Indicators of National Development, in: Policy Sciences, 4 (1973), pp. 63–74.

<sup>14</sup> Galtung, Johan, et al., Measuring World Development, in: Alternatives, 1 (1975), pp. 131–158, 523–555.



try, the similarities have been outstanding when compared with variations in what is meant by political development.

Some international civil servants and scholars seem to think that international official statistics might include more specifically political indicators in the future. Governments may become more responsive to requests for data as we have more specific notions of what we want. At least the probability is great enough that we need to think carefully about the question.

Several efforts have already been undertaken, of course. A vast amount of political information, from which quantitative series can be derived, is published annually in such publications as the *Statesman's Yearbook*, the *Europa Year Book* and the *Political Handbook of the World*<sup>15</sup>. Books such as the *World Handbook of Political and Social Indicators* go further in attempting to produce quantitative series<sup>16</sup>. A much larger number of efforts are made to measure a more limited scope of political activity. Freedom House, for example, publishes annually indicators of the status of human rights around the world<sup>17</sup>. In a different area, both the International Institute for Strategic Studies and the Stockholm International Peace Research Institute publish data annually on allocations of money and manpower to the defense sector by governments. These and similar data have been used by Sivard to demonstrate the degree to which the world's resources are used for military rather than peaceful uses<sup>18</sup>.

Many sources that are available can be employed to produce measurements of government structures and changes in those structures, allocations of government resources, freedom and repression, and political participation. Comparative measures of the nature and functions of legislatures, for example, can be developed from material available in a number of publications. Does the executive have a veto over legislation or not? Where does legislation originate? Is the executive responsible in any way to the legislature? Is there effective party discipline or not? Is, in fact, the legislature divided by parties? Whether or not this is so, is opposition allowed? Is the government formed from one party or from a coalition? Is it represented directly in the legislature or is it separated from it by constitutional structures? Similar questions can, of course, be posed for the executive, judicial and administrative parts of the governmental structure and for the party system. Changes in these structures and changes of personnel within them can be monitored by newspaper and other current accounts of events.

<sup>15</sup> Paxton, John (ed.), *The Statesman Year-Book*, London, annual; *The Europa Year-Book 1977: A World Survey*, London 1977; Banks, Arthur S. (ed.), *Political Handbook and Atlas of the World: Parliaments, Parties and Press*, Council on Foreign Relations (ed.), New York, annual.

<sup>16</sup> Taylor, Charles Lewis, and Hudson, Michael C., *World Handbook of Political and Social Indicators*, 2nd. ed., New Haven/Conn. 1972.

<sup>17</sup> Gastil, Raymond D., *The Comparative Survey of Freedom*, in: *Freedom at Issue* (1973–1979).

<sup>18</sup> Sivard, R. L., *World Military and Social Expenditures*, Leesville/Virginia 1977.

Another set of structural questions relate to center-periphery relationships. To what extent are decisions of various kinds made in the central government and to what extent is there leeway for changes or independent decisions at the provincial or local levels? Are the localities homogeneous with the central government? Is there internal colonialism and subjected peoples within the central government? Not all imperialism need necessarily stretch across water. If ethnically or otherwise identifiably different people are present within a state, how integrated into the political system are they? To what extent are they represented in leadership and other decision making roles?

Ted Gurr in reviewing the *World Handbook of Political and Social Indicators: Second Edition* (1972) suggested that further collections ought to get to the underside of non-official life<sup>19</sup>. How is it with the powerless? The incidence of protests, at least when it reached public proportions, was reported but much more specific data can be codified as to who the dissidents are, what they want and how much (or little) they already have. Of course, participation of several sorts may be measured. Conventional activity such as voting and even group membership is relatively easy to collect. Less conventional kinds may be somewhat more difficult without public opinion surveys. While these are being conducted in an increasing number of countries, they probably cannot be properly considered social bookkeeping sources.

Closely related to participation is the question of freedom. Is there freedom for groups and individuals to express themselves and to oppose the government? Is there political oppression? Again present social bookkeeping sources do not include very many data of this kind. Finding comparative data from government sources is unlikely to be possible in any case in the real world. Nevertheless, this important aspect of political life needs to be considered carefully for ways of measuring it.

Data on the public versus private (or semi-private) sector, on productivity and investment in each, on allocations to education, health and welfare, on personnel deployment and a number of other areas of political life might be added as examples of kinds of quantitative information that either can be derived from current social bookkeeping sources or that should be encouraged on governments and official statisticians for collecting.

It would be difficult to call all of this a system of political indicators. It is more at the moment a hodgepodge of ideas. Political scientists have yet some way to go before arriving at even as much consensus on political indicators as sociologists have achieved on social indicators. Cross national comparisons of political systems have been based primarily on non-political variables or have been limited to a few countries and based upon qualitative data. The first step toward more rigorous comparisons for many countries on the basis of political indicators is, therefore, the careful development of these indicators. From this eventually might arise a system and perhaps even a generally accepted model.

<sup>19</sup> Gutt, Ted Robert, *The Neo-Alexandrians: A Review Essay on Data Handbooks in Political Science*, in: *American Political Science Review*, 68 (1974), pp. 243–252.

#### 4. Conclusion

Data systems are designed to give to particular users, particular kinds of analyses, and therefore particular kinds of results. This is not to suggest that data systems are means for scholastic cheating; it is rather to insist that theory building has already begun with the choices required in the measurement of data. In the social sciences, the data system most fully developed for cross national comparison is that of the National Account Statistics. While many criticisms have been levelled at it for what it does not do, it does what it does well. It appears to be amenable to adaption for other purposes as well. One of these purposes is the study of well-being or general welfare in social life. The Social Indicators Movement, however, has not limited itself to revisions of national accounts but is seeking a variety of approaches to a social measurement system. For cross national quantitative political comparisons, virtually no systematic designs have been undertaken except for political violence studies. This work is still in the stage of definition. It clearly needs more useful political indicators.

The Analysis of Change and Persistence in German Society:  
The German Census of Occupations as a New Data Base

0. Foreword

Recently historians and sociologists have taken increased interest in the study of German society during the last two centuries<sup>1</sup>. But quantitative analyses of societal developments in Germany in the nineteenth and twentieth century are quite rare. This is the more astonishing as there is a data base which offers a number of possibilities for quantitative analyses: the German census of occupations. Not the least of the reasons for lack of interest in the German census of occupations are the problems confronting the handling of the vast statistical tables of the published census volumes. The purpose of this article is to illustrate the scope of the German census of occupations, to point out to its problems and to formulate some questions which were not raised up to now and which could be answered by analysing the German census data. The argument will be that a preparation of these data for machine-readable use will create a new data base which allows new research questions.

<sup>1</sup> Studies which were carried out before 1976 are recorded in: Wehler, Hans-Ulrich, *Bibliographie zur modernen deutschen Sozialgeschichte*, and in: Wehler, Hans-Ulrich, *Bibliographie zur modernen deutschen Wirtschaftsgeschichte*, both Göttingen 1976. The Study of German Society in the past has gained a lot of interest from quantitative historians in Germany. This is one of the results of the survey, QUANTUM carried out in the Federal Republic of Germany, Austria and the German-speaking parts of Switzerland. See: Bick, Wolfgang, et al., QUANTUM DOKUMENTATION. *Quantitative Geschichtswissenschaft 1977/Quantitative History 1977*, Stuttgart 1977, and by the same authors: *Quantitative History in Transition*, *Social Science Information*, 16 (1976), pp. 679–714.

## 1. Background

Among German historians, the census of occupations is a rarely used statistical source for describing German economy and society in the past<sup>2</sup>. The usual approach has been the attempt to describe change in the occupational structure as a part of social change in Germany<sup>3</sup>. Mostly, these studies cover only one geographical and/or political/administrative unit: entire Germany, or entire German states, e. g. Prussia, Bavaria, etc., or smaller regions. Comparisons across unit-boundaries are neglected. Besides, these studies often restrict themselves to analyses of specific dates or short periods of time in the past. Research that deals with longer time spans is rare as well, and, if carried out, uses only selected information for time comparisons. This use of data from the census of occupations, either focussing on the nation state and/or neglecting comparisons across boundaries, resulted on a research bias, which left out important within-nation variations. It also led to misinterpretations due to lack of analyses over long periods of time. Preparation of the census data for machine-readable use will permit systematic comparisons across boundaries and over time. Moreover the census data processed into machine-readable form will permit cross-tabulation of variables which are not now available for sophisticated analyses and which cannot be used for answering elaborated research questions.

## 2. The History of the German Census of Occupations

The first comprehensive German census of occupations covering the entire German population was carried out in 1882<sup>4</sup>. Earlier occupational surveys were carried out by the German states only. Although German occupational surveys are known since the eighteenth century<sup>5</sup>, only nineteenth- and twentieth-century data are preserved

<sup>2</sup> Gustav Schmoller was one of the first to use data from occupational censuses: Schmoller, Gustav, *Zur Geschichte der deutschen Kleingewerbe im 19. Jahrhundert*, Halle 1870. Among sociologists, the most prominent example for the use of census data is Theodor Geiger's analysis of the 1925 occupational census: Geiger, Theodor, *Die soziale Schichtung der Deutschen Volkes*, Stuttgart 1932.

<sup>3</sup> See the compilation of imperial statistics by Hohorst, Gerd, et al., *Sozialgeschichtliches Arbeitsbuch. Materialien zur Statistik des Kaiserreiches 1870–1914*, München 1975.

<sup>4</sup> See: Fürst, Gerhard, *Wandlungen im Programm und in den Aufgaben der amtlichen Statistik in den letzten 100 Jahren*, in: *Statistisches Bundesamt (ed.), Bevölkerung und Wirtschaft 1872–1972*, Stuttgart 1972, pp. 13–82.

<sup>5</sup> For a history of the pre-1860 Prussian Statistics see: Boeckh, Richard, *Die geschichtliche*

and accessible in a way that allows their preparation for machine analysis. But as almost all of the pre-1882 surveys were carried out for administrative purposes only, the recorded informations were partially published, mostly in publications of statisticians from the state statistical offices that carried out the surveys or compiled the data from the conducting institutions, the local state authorities. Although no extensive archive research has been carried out as yet, it is to be assumed that most of the manuscript schedules have been destroyed, either by the authorities themselves or during the Second World War. There are some manuscript schedules in local archives which have been used for research, for example from the Rhineland area<sup>6</sup>, but the number of preserved manuscript schedules seems to be too small for preparation of a data file covering more than a few cities and counties. This means that a machine-readable census-data file for the pre-1882 censuses of occupations has to be based on published census material, if no further census material can be discovered by intensive archive investigations.

Before the foundation of the German empire in 1871 the results of occupational surveys have been published separately by a number of German States. For the introductory purpose of this paper, only the comprehensive published tables of the Prussian Censuses of occupations from the census years 1849, 1855, and 1858 will be dealt with.

The first census of occupations which covered the entire German population was carried out in 1882. Subsequent censuses of occupations were carried out in 1895, 1907, 1921, 1933, and 1939. The results of each year's census was published in a number of folio-volumes for each census year<sup>7</sup>. The survey and the publication of the results usually involved considerable work: For the 1895 census of occupations (and of industrial establishments) 41.6 million questionnaires were distributed, 1600 people were employed for compiling the data and for calculating the tables, and it is estimated that, including helping family members of the employees, some 6500 people took part in the preparation of the data for publication.

Entwicklung der amtlichen Statistik des preußischen Staates, Berlin 1863. Further information on Prussian statistics of this period is given by: Kaufhold, Karl-Heinrich, Inhalt und Probleme einer preußischen Gewerbestatistik von 1860, in: Wirtschaftliche und soziale Strukturen im säkulareren Wandel. Festschrift für Wilhelm Abel zum 70. Geburtstag, Bog, Ingomar, et al. (eds.), Vol. 3, Hannover 1976, pp. 707–719.

<sup>6</sup> See for example: Hoth, Wolfgang, Die Industrialisierung einer rheinischen Gewerbestadt – Dargestellt am Beispiel Wuppertal, Köln 1975.

<sup>7</sup> Rauchberg, Heinrich, Die Berufs- und Gewerbezahlung im Deutschen Reich vom 14. Juni 1895, Berlin 1901, p. 29.

### 3. The Prussian Census of Occupations 1849, 1855 and 1858<sup>8</sup>

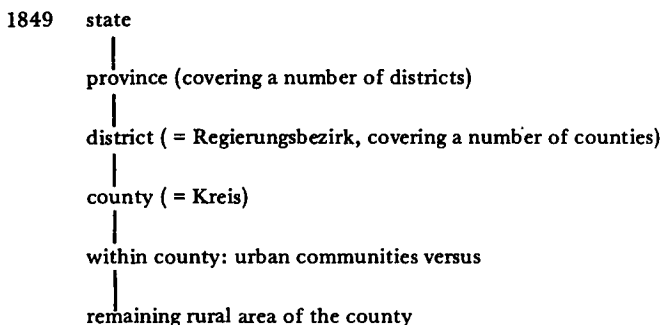
The published tables show the number of masters, journeymen, apprentices, and assistants for recorded occupation within specific administrative units. The number of occupations recorded was in

1849: 243<sup>9</sup>

1855: 271

1858: 325

The figures for each occupation were aggregated on different levels:



The system of occupational categories, used in the Prussian censuses of occupations of the first half of the nineteenth century was also used in the censuses of 1882, 1895, and 1907 census. This system, which was heavily discussed in statistical literature of that time<sup>10</sup>, was abolished with the 1925 census. In theory the classifica-

<sup>8</sup> — Tabellen und amtliche Nachrichten über den preußischen Staat für das Jahr 1849. Vol. 5: Die Gewerbe-Tabelle enthaltend: Die mechanischen Künstler und Handwerker, bei denen der Meister mit Gehülfen arbeitet, die Anstalten und Unternehmungen zum literarischen Verkehr gehörig, die Handelsgewerbe, die Schifffahrt, das Fracht- und Lohnfuhrwesen, die Gast- und Schankwirtschaft, die Civilbeamten in Staatsdiensten, die Kommunalbeamten, die ländlichen Erwerbs-Verhältnisse sowie die Handarbeiter und das Gesinde, für 1849 und 1852, Berlin 1854; — Tabellen und amtliche Nachrichten über den preußischen Staat für das Jahr 1855, Berlin 1858; — Tabellen und amtliche Nachrichten über den preußischen Staat für das Jahr 1858, Berlin 1860.

<sup>9</sup> A table from the 1849 census of occupations is displayed in Appendix I.

<sup>10</sup> Rauchberg, Berufszählung, pp. 1–18; Losch, H., Berufsprobleme und berufsstatistische Probleme, in: Württembergische Jahrbücher für Statistik und Landeskunde (1919/20), pp. 259–267; Meerwarth, Rudolf, Nationalökonomie und Statistik, Berlin und Leipzig 1925, pp. 28–108; Fürst, Gerhard, Zur Methode der deutschen Berufsstatistik, Allgemeines Statistisches Archiv, 19 (1929), pp. 1–29.

tion system used before the First World War was based on „pure“ occupational categories, but in reality it was a mixed system: Some occupations were denominated „purely“ occupationally (e. g. baker, blacksmith, etc.), some were categories of goods and articles specific occupations sold or produced (e. g. millinery), and some denominated the business establishments, certain occupations held or ran (e. g. mill, pharmacy, etc.).

#### 4. The Imperial Census of Occupations 1882 – 1895 – 1908 – 1925 – 1933 – 1939<sup>11</sup>

The picture shown by the imperial census of occupations is quite different from that one gets from the early Prussian census. The number of occupations recorded by the surveys increased dramatically

1882: more than 6000 occupations  
1895: more than 7000 occupations  
1907: more than 8000 occupations  
1925: not yet known  
1933: not yet known  
1939: not yet known

Before entering the published statistical tables, the mass of recorded occupations was grouped according to specific classification rules. The classification system was based, firstly, on broad economic sectors („Berufsabteilungen“) and, secondly, subsumed occupational classes („Berufsarten“) under these „Berufsabteilungen“.

<sup>11</sup> — Berufsstatistik nach der allgemeinen Berufszählung vom 5. Juni 1882, Statistik des deutschen Reiches, N. F., Vols. 2–4;  
— Berufs- und Gewerbebezahlung vom 14. Juni 1895, Statistik des Deutschen Reiches, N.F., Vols. 102–111 (census of occupations only);  
— Berufs- und Betriebsstatistik vom 12. Juni 1907 – Berufsstatistik, Statistik des Deutschen Reiches, Vols. 202–211;  
— Berufszählung. Die berufliche und soziale Gliederung der Bevölkerung des Deutschen Reiches (1925 census of occupations), Statistik des Deutschen Reiches, Vols. 402–408;  
— Die berufliche und soziale Gliederung der Bevölkerung in den Ländern und Landesteilen. Volks-, Berufs- und Betriebszählung vom 16. 6. 1933, Statistik des Deutschen Reiches, Vols. 455–457;  
— Volks-, Berufs- und Betriebszählung vom 17. 5. 1939, Statistik des Deutschen Reiches, Vols. 555–557.



The „Berufsabteilungen“ for the 1895 census were:

- A. Agriculture
- B. Mining, Building, Industry, Different Branches of Industry
- C. Trade
- D. Unskilled Labor and Domestic Services
- E. Civil and Military Services and (=including) Professions
- F. Without Occupation

According to the lines of theses „Berufsabteilungen“, some thousand occupations were subsumed under a few hundred occupational classes (e.g. 207 occupational classes for the 1895 census) in the following way<sup>12</sup>:

<i>A. Agriculture</i>	<i>C. Trade</i>
Occupational Class („Berufsart“) A1	Occupational Class C1
Occupational Class A2	...
...	...
...	Occupational Class C <sub>n</sub>
Occupational Class A <sub>n</sub>	
	<i>D. ....</i>
<i>B. Mining etc.</i>	...
Occupational Class B1	...
...	
...	<i>E. ....</i>
Occupational Class B <sub>n</sub>	...
	...
	<i>F. ....</i>
	...
	...

On the first fifty pages of the volumes of each census year, the coding scheme for recording occupations is given.

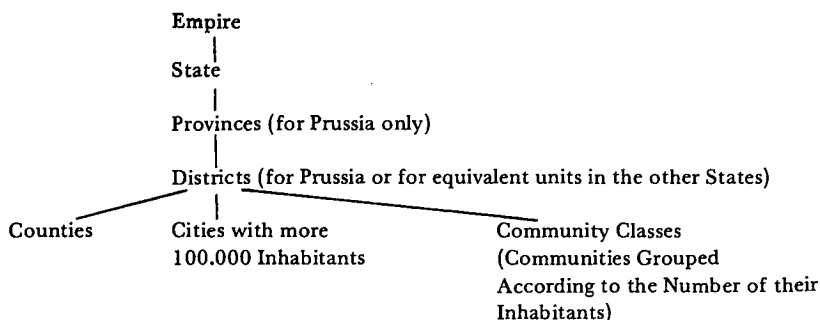
Data for the following variables were entered into the published tables for recorded „Berufsabteilung“ and for recorded „Berufsart“:

- |                          |  |
|--------------------------|--|
| 1. Chief Occupation      | 6. Age Groupings                       |
| 2. Additional Occupation | 7. Employees of People Listed under 1. |
| 3. Occupational Status   | 8. Without Occupation                  |
| 4. Personal Status       | 9. (Since 1907) Place of Birth         |
| 5. Sex                   |  |

This scheme changed somewhat from the 1882 census to the 1939 census, but its general structure was kept unchanged until 1939.

<sup>12</sup> A page from the 1895-coding-scheme is displayed in Appendix II.

Information for each „Berufsabteilung“ and „Berufsart“ was entered into the tables for the following geographical and/or political/administrative units (i. e. levels of aggregation):



Information for the complete list of variables were displayed for the higher aggregation levels only, on the county level only the data for variables 1 (chief occupation) and 2 (additional occupation) was given in the published tables.

## 5. Problems

### 5.1 Compatibility Problems

As mentioned above, the rules for classifying occupations were changed for the 1925 and later censuses of occupations. The classification system used from 1882 to 1907 grouped single occupations according to broader occupational groups („Berufszugehörigkeit“). Although there were deviations from this principle – occupations were grouped according to goods sold or produced or according to business or industrial establishments individuals ran or were attached to – occupations were in general grouped along the lines of occupational activities. From the 1925 census on, occupations were grouped primarily according to types of industrial or business establishments individuals were attached to („Betriebszugehörigkeit“)<sup>13</sup>. This involved substantial alterations of the classification system, so that a file of continuous indicators of change across the 1907–1925 line cannot be established. It is to be

<sup>13</sup> Fürst, Methode, pp. 9–13.

assumed that the same applies to an incorporation of occupational census data from the pre-imperial period. The compatibility problems involved in a continuous data file from the first half of the nineteenth century to 1907 allows only the possibility of establishing two files for the period concerned.

### 5.2 Boundary Problems

The establishment of a continuous data file for the entire period from the first half of the nineteenth century to 1939 is further aggravated by changes of the boundaries of the units for which data were recorded. Although results of intensive research on boundary changes within Germany are not yet available, it appears — after a *prima facie* survey of the problems involved<sup>14</sup> — that major boundary changes took place in the sixties and seventies of the nineteenth century and between 1910 and 1920.

These boundary problems and the compatibility problems indicate the necessity of establishing three files of continuous census data instead of one file. The establishment of a data file for the first half of the nineteenth century, a second file for the years 1882–1907, and a third file for the years 1925 to 1939.

## 6. The Census of Occupations as a New Data Base for Answering Questions about German Society in the Past

Two possibilities of analysis that a machine-readable set of occupational census data will allow, have been mentioned already: time series analysis and comparisons across boundaries. Besides that, research questions about society in German history that have only been sketched could be answered. The notion of Germany as a nation which was dominated in its past by sharp regional differentiation — a notion often postulated but rarely investigated<sup>15</sup> — will be the focus of research based on occupational data. Aspects of a social ecology which have important implications for de-

<sup>14</sup> During the period concerned all political/administrative units changed.

<sup>15</sup> Among historians, especially scholars from economic history recognize this problem. See: Tipton Jr., Frank B., *Regional Variations of Germany during the Nineteenth Century*, Middleton/Conn. 1976.

termining personal experiences and for influencing individual behaviour could be analyzed<sup>16</sup>.

The more specific questions that could be answered by data from the censuses of occupations as a machine-readable data file are about

- the differentiation of the occupational structure in general;
- the concentration of occupational classes in urban or in rural areas or in certain states or regions;
- the configurations of primary and additional occupations;
- occupations and mean family size;
- female labor.

This list is of course incomplete and does not elaborate theoretical interconnections. Its purpose is that of this entire article: to find paths through a vast research field. Further efforts, hopefully in cooperation with others who puzzle over the same problems will follow to make these paths into broad avenues.

<sup>16</sup> Among sociologists Erwin K. Scheuch and Ralf Dahrendorf pointed out to these problems: Scheuch, Erwin K., *Continuity and Change in German Social Structure*, in: Kissinger, Henry A. (ed.), *In Search of Germany*, mimeo., Cambridge/Mass., Harvard Center for International Affairs 1966, pp. 82–207 and 762–818. Dahrendorf, Ralf, *Gesellschaft und Demokratie in Deutschland*, München 1968.

Appendix I: Table from the 1849 Prussian Census of Occupations

Namen der Kreise.		Städte und plattes Land.	Mechanische Künstler und Handwerker.									
			Steinsetzer oder Plünsterer.		Schornsteinfeger.		Töpfer und Ofen- Fabrikanten, auch Verfertiger von irdenem Geschirr.		Glaser und Glaschleifer.		Bilder-, Blumen- Porzellanmaler, Daguerretypisten und Koloristen.	
			Meister, oder für eigene Rechnung arbeitende Personen.	Helfen und Lehrlinge.	Zahl des für eigene Rechnung arbeitenden Personen.	Helfen und Lehrlinge.	Meister, oder für eigene Rechnung arbeitende Personen.	Helfen und Lehrlinge.	Meister, oder für eigene Rechnung arbeitende Personen.	Helfen und Lehrlinge.	Zahl des für eigene Rechnung arbeitenden Personen.	Helfen und Lehrlinge.
XV. Regierungsbezirk Magdeburg.												
		75.	76.	77.	78.	79.	80.	81.	82.	83.	84.	
Kalbe.	Schönebeck . . .	1	2	1	3	3	5	6	1	"	"	
	Kalbe a.d.Saale . . .	"	"	1	1	2	4	4	2	"	"	
	Aken . . . . .	"	"	"	"	2	1	2	"	"	"	
	Barby . . . . .	1	2	1	1	4	7	2	1	"	"	
	Gr. Salze . . . . .	2	11	"	"	1	1	2	"	"	"	
	Stassfurt . . . . .	"	"	1	"	"	"	1	1	"	"	
	Plattes Land . . .	"	4	"	"	2	2	2	"	"	"	
Summe		4	19	4	5	14	20	19	5	"	"	
Wanzleben.	Egeln . . . . .	"	"	1	1	5	7	1	"	"	"	
	Wanzleben . . . . .	1	3	1	2	"	"	3	"	1	"	
	Seehausen . . . . .	"	"	"	"	3	1	3	"	"	"	
	Halmersleben . . . . .	"	"	"	"	"	"	1	"	"	"	
	Plattes Land . . .	2	14	1	1	1	"	2	1	"	"	
Summe		3	17	3	4	9	8	10	1	1	"	
Magdeburg.	Magdeburg . . . . .	6	26	3	13	6	11	34	31	12	3	
	Vorstadt Neu- stadt . . . . .	1	18	"	"	1	"	3	3	1	"	
	Vorstadt Sü- denburg . . . . .	"	1	"	"	"	"	1	1	1	9	
	Summe	7	45	3	13	7	11	38	35	14	12	
Wollmirstädt.	Wollmirstädt . . .	"	"	2	4	1	1	3	2	"	"	
	Plattes Land . . .	1	6	"	"	"	"	2	"	"	"	
	Summe	1	6	2	4	1	1	5	2	"	"	
Oschersleben.	Oschersleben . . .	"	"	2	2	1	3	5	2	"	"	
	Wegeleben . . . . .	"	"	"	"	"	"	2	"	"	"	
	Gröningen . . . . .	"	"	1	"	"	"	2	"	"	"	
	Kroppenstädt . . .	1	3	"	"	"	"	1	"	"	"	
	Schwanebeck . . .	"	"	"	"	1	"	1	"	"	"	
	Plattes Land . . .	"	"	1	"	1	1	2	"	"	"	
Summe		1	3	4	2	2	4	13	2	"	"	
Aschersleben.	Quedlinburg . . .	1	5	3	4	1	"	6	1	1	"	
	Aschersleben . . .	1	5	1	2	"	"	4	2	2	"	
	Kochstädt . . . . .	"	"	1	1	1	"	"	"	"	"	
	Plattes Land . . .	"	"	"	"	"	"	2	"	"	"	
Summe		2	10	5	7	2	"	12	3	3	"	
Wernigerode.	Wernigerode . . .	2	2	1	3	2	2	5	1	"	"	
	Plattes Land . . .	1	2	"	"	"	"	4	"	"	"	
	Summe	3	4	1	3	2	2	9	1	"	"	
Halberstadt.	Halberstadt . . . . .	1	21	1	4	3	2	9	5	4	"	
	Ostervick . . . . .	"	5	2	"	"	"	3	1	"	"	
	Hornburg . . . . .	"	"	"	"	3	3	3	"	1	"	
	Derenburg . . . . .	"	"	1	1	"	"	1	"	"	"	
	Dardeshelm . . . . .	"	"	1	1	1	1	2	"	"	"	
	Plattes Land . . .	2	18	"	"	"	"	2	"	"	"	
Summe		3	44	5	6	7	6	20	6	5	"	
In den 49 Städten . . . . .		27	124	46	70	101	108	199	72	26	15	
In den 14 Kreisen des platten Landes . . . . .		7	57	7	3	29	36	53	6	3	7	
Im ganzen Regierungsbezirke		34	181	53	73	130	144	252	78	29	22	
XVI. Regierungsbezirk Merseburg.												
Delitsch.	Eilenburg . . . . .	2	3	2	5	4	6	11	7	"	"	
	Delitsch . . . . .	1	1	1	2	2	3	4	3	"	"	
	Landberg . . . . .	"	1	1	1	"	"	2	1	"	"	
	Plattes Land . . .	1	1	"	"	4	4	4	1	"	"	
Summe		4	6	4	8	10	13	21	12	"	"	

# Mechanische Künstler und Handwerker.

Zimmer-, Schilder-, Roulexmaler, Anstreicher, Vergol- der, Stäfler, Stuk- kateur, Goldleisten- und Goldrahmen- macher.		Lackirer aller Art, als: Blech-, Holz-, Tuch- etc.		Bildhauer, Kupfer- stecher, Formstecher, Furnaceinnehmer und Ciseleure.		Grobschmiede aller Art, als: Groß-, Halb-, Wullen-, Pfannen-, Ketten-, Senzen- etc. Schmiede.		Schlosser, worunter auch Zirkel-, Zeug-, Bohr-, Säge-, Messer-, Nagel-, Büchsen- schmiede, Sporer, Feilenbauer, Instru- mentenschleifer, Schweireuschleifer und Maschinenbauer.		Gürtler-, Braceur-, Neugold-, Neu- silber-Arbeiter, Metallknopfmacher und Schwerdtfeiger.		Mühlbauer und Mühlendick- Arbeiter.	
Zahl der für eigene Rechnung arbeitenden Personen.	Gehülten und Lehrlinge.	Zahl der für eigene Rechnung arbeitenden Personen.	Gehülten und Lehrlinge.	Zahl der für eigene Rechnung arbeitenden Personen.	Gehülten und Lehrlinge.	Meister, oder für eigene Rechnung arbeitende Personen.	Gehülten und Lehrlinge.	Meister, oder für eigene Rechnung arbeitende Personen.	Gehülten und Lehrlinge.	Meister, oder für eigene Rechnung arbeitende Personen.	Gehülten und Lehrlinge.	Zahl der für eigene Rechnung arbeitenden Personen.	Gehülten und Lehrlinge.
85.	86.	87.	88.	89.	90.	91.	92.	93.	94.	95.	96.	97.	98.
1	2	»	»	»	»	6	14	18	23	»	»	»	»
2	»	»	»	»	»	5	17	17	20	»	»	»	»
»	»	»	»	»	»	5	6	5	7	»	»	»	»
1	»	»	»	»	»	5	6	4	2	»	»	»	»
1	1	»	»	»	»	3	3	4	2	»	»	»	»
1	1	»	»	»	»	6	7	5	6	»	»	»	»
»	»	»	»	»	»	49	42	7	7	»	»	»	»
5	3	»	»	»	1	79	84	63	67	»	»	»	»
1	2	»	»	»	»	3	6	8	9	1	»	»	»
3	»	»	»	»	»	3	4	5	6	»	»	»	»
»	»	»	»	1	»	5	1	4	5	»	»	1	»
»	»	»	»	»	»	1	»	2	3	»	»	»	»
»	»	»	»	»	»	72	86	12	14	»	»	3	9
4	2	»	»	1	»	84	97	31	37	1	»	4	9
58	52	18	20	6	8	22	55	87	178	10	7	1	»
3	»	»	»	»	»	8	18	11	26	»	»	1	2
4	4	»	»	»	»	5	25	8	49	»	»	»	»
65	56	18	20	6	8	35	98	106	253	10	7	2	2
2	2	»	»	»	»	4	7	5	8	»	»	»	2
2	2	»	»	»	»	72	74	8	7	»	»	»	2
2	2	»	»	»	»	76	81	13	15	»	»	»	2
2	4	»	»	»	»	4	6	6	7	1	»	1	»
»	»	»	»	»	»	4	4	3	1	»	»	»	»
»	»	»	»	»	»	4	7	3	2	»	»	»	»
»	»	»	»	»	»	5	2	2	»	»	»	»	»
»	»	»	»	»	»	4	3	3	1	»	»	»	»
»	»	»	»	»	»	41	35	3	3	»	»	»	»
2	4	»	»	»	»	62	57	20	14	1	»	1	»
6	2	4	»	»	»	17	23	26	25	4	3	»	»
6	1	»	»	»	»	13	12	24	21	2	»	1	1
»	»	»	»	»	»	3	1	1	»	»	»	»	»
»	»	1	»	»	»	41	27	6	»	»	»	1	2
12	3	4	1	»	»	74	63	57	50	6	4	2	3
4	4	1	2	»	»	9	9	7	7	1	»	1	»
4	1	»	»	1	1	24	18	11	17	1	»	»	»
4	5	1	2	1	1	33	27	18	24	2	»	1	»
18	8	3	4	1	2	18	35	34	25	5	3	1	3
1	»	»	»	»	»	3	4	6	7	»	»	»	»
1	»	»	»	»	»	3	4	7	2	»	»	»	»
»	»	»	»	»	»	3	3	1	1	»	»	»	»
»	»	»	»	»	»	4	1	2	»	»	»	»	»
»	»	»	»	»	»	52	29	2	2	»	»	»	»
20	8	3	4	1	2	83	73	52	37	5	3	1	3
155	104	26	26	11	10	281	408	558	651	31	19	16	18
5	5	1	1	3	1	926	744	100	86	1	»	9	14
160	109	27	27	14	11	1,207	1,152	658	737	32	19	25	32
2	1	»	»	»	»	8	15	30	41	4	3	»	1
1	»	»	»	»	»	7	4	21	30	4	3	»	»
»	»	»	»	»	»	3	3	5	5	»	»	»	»
»	»	»	»	»	»	119	65	3	2	»	»	»	12
3	1	»	»	»	»	137	87	59	78	8	6	»	13

36\*  
(B 20

Fingerringe, goldene, 3  
Goldarbeiter  
Goldfärber  
Goldketten f  
Goldpolirer  
Goldprobirer  
Gold- und Silberarbeiter

Gold- und Silberschmiede  
Goldwaaren V  
Goldwaarenfitter  
Goldwaarenputzer  
Granatschmuck V  
Juwelenarbeiter  
Juwelenfasser

Juweliere  
Knopf F (in Gold u. Silber)  
Kreuze, goldene u. silberne,  
B f  
Manschetten- und Hemden-  
Knöpfe, goldene und  
silberne. B f

Medaillons und Broschen,  
goldene und silberne. B f  
Ohrringe, goldene. B f  
Orden f  
Ring f (in Gold u. Silber)  
Schmuckwaaren, goldene und  
silberne. B f

Silberarbeiter  
Silberpolirer  
Silberwaaren &  
Steinfasser  
Tafelgeräthe &

## B21 Sonstige Verarbeitung edler Metalle (Gold- und Silberschläger, Drahtzieher), Münzstätten und Prägeanstalten

Agraffen F  
 Bijouteriewaaren, goldene  
 und silberne, B f  
 Blattgold und Blattsilber f  
 Profatwaaren F  
 Büchlein B  
 Chatons, goldene und  
 silberne, B f  
 Doubléwaaren F  
 Drahtspinner (für Gold  
 und Silber)  
 Drahtzieher (f. Gold und  
 Silber)  
 Edelmetallpräger  
 Feingoldschläger  
 Feilgranarbeiter

Flügelstischläger  
Flüsterkasseler  
Galaxie, goldene u. silb., 20  
Gefäßmacher  
Gefäßschmelzer  
Goldbortenmacher für ge-  
wirkte Waaren  
Goldbühnenmacher  
Goldbraut &  
Goldbrautspinnern, -zieher  
Goldbrautwaaren &  
Goldgepinns &  
Goldglätter  
Goldklammer  
Goldplattir  
Goldpreller

Goldschläger  
 Gold- und Silberdrahtzieher  
 Gold- und Silber-Gefäß-  
 schmelzereien  
 Gold- und Silber-Plättlein-  
 schläger  
 Gold- und Silberbleisler  
 Gold- und Silberpinnenreier  
 und •Weberer  
 Gold- und Silberstücker  
 Gold- u. Silberwalzgeschäfte  
 Goldwäcker  
 Kettengeschäfte  
 Ketten, Kettenbeschläge, Ge-  
 hänge u., goldene und  
 silberne. 34

Rührschmelzer  
 Rührgeschäfte  
 Rührmühlen  
 Leonische Waaren, B f  
 Medaillenprägenstalten (für  
 Medaillen aus edlen Me-  
 tallen)  
 Metallschläger f. edle Metalle  
 Münzpräger,  
 Münzstätten  
 Plättleinschläger  
 Prägenstalten für edle Me-  
 talle  
 Silberbortenmacher  
 Silberdruck f

Silberdrahtspinner, -zieher  
Silberdrahtwaaren f  
Silberpresser  
Silberschläger  
Silberspinner  
Silberstricker  
Silberwirker  
Treffen f  
Zulawaaren f  
Uhrbügel, Uhrhaken,  
ketten, Uhrschlüssel,  
dene und silberne, v  
Zergoldder und Versilber:

## B 22 Aufschmiede.

**Reßeländer  
Reßler**

### Rupfer schläger

## Rupferſchmiede

Rupferwaaren, grobe, B f

Roßhymie

B 23. Blut- und Geisgießer.

Gelbgießer

Messinggewichte 8

**Messinggießer**

Roßgießer

## B 24 Zinngräber.

Rannengießer  
Löffelgießer (=Präger)  
Rinndrehler

Binnfiguren &  
Binngießer

Binnfurgwaaren B  
Binnlichtformen F

Zinnlöffel F  
Zinnpfund F

Binnſpreken 8  
Binnwaaren 8

### B 25 Herstellung von Spielwaren aus Metall.

Blechspielwaaren &  
Eis- und Zinnspielwaaren &

Kinderfädel B  
Kindertrommel K

Spielwaren aus Metall, B f

Stimmacher

## Binnspielwaren &

**B 36 Sonstige Verarbeitung unedler Metalle, mit Ausnahme von Eisen.**

Aluminiumwaaren ♂ (mit  
 Ausnahme von Schreib-  
 federn)  
 Blattmetall ♂  
 Bleidraht ♀  
 Bleispielen ♂  
 Bleischießohr ♀  
 Bleigießer  
 Bleitungel ♂  
 Bleitungelpresser  
 Bleiplomben ♂  
 Bleischrot ♂  
 Blei- und Zinnrohr ♂  
 Bleinwaaren ♂  
 Bleigießer  
 Drahtspinner für Kupfer  
 Drahtzieher für Kupfer

Fensterblei F  
 Füllter F  
 Galbanifeyre  
 Galbanoplastische Anstalten  
 Galbanoplastische Waaren W  
 Gelschmelbemaßer  
 Glanzgoldmacher  
 Glanzroß F  
 Hagel F  
 Kugel F  
 Kugelpresser  
 Kupferdrahtspinner, -zieher  
 Kupferwaaren, feine, B f  
 Radierer W  
 Metallarbeiter  
 Metalldraht W  
 Metalldreher

Metallgießer  
 Metallhammerwerke  
 Metallkapsel F  
 Metallplastiker  
 Metallpolirer  
 Metallschablonen F  
 Metallschmelzer  
 Metallschneider  
 Nickelblech F  
 Nickelketten F  
 Nickelwaaren F  
 Patronenhülsen F  
 Pfeifenbeschlag  
 Pfeifendeckelmader  
 Plaquewaaren F  
 Plattirer

Plattirwaaren f  
 Plouben f  
 Plombengießcr  
 Prägenaff (außgenom-  
 men für edle Metalle)  
 Schrot f  
 Silberimitationswaaren f  
 Silberplattir  
 Spialler  
 Sporenmacher  
 Sporer  
 Staniel f  
 Stanielöfen f  
 Stanielkapsel B  
 Stanielschläger  
 Tuben B

Verniedelungskunstalten  
 Verzinker  
 Verzinner  
 Zindbraht  $\mathfrak{B}$   
 Zinggießer  
 Zinformament  $\mathfrak{B}$   
 Zinsprägenstollen  
 Zinsringel  $\mathfrak{F}$   
 Zinkwaaren  $\mathfrak{F}$   
 Zindbraht  $\mathfrak{B}$   
 Zinnfolien  $\mathfrak{F}$   
 Zinnhülsen  $\mathfrak{F}$   
 (farben)  
 Zinnladierer  
 Zinnmaler  
 Zinnstücker

## B 27 Erzgießer, Glockengießer.

Bildgießerei  
Bronzegießer

Eraggießer

Blodengießer

### Glodenreparateur

**Glockenschleifer**

## Standardization of Longitudinal, Aggregate-Level Data in the Norwegian Commune Database

### Introduction

In Norway, communes are the main aggregate-level units for the collection of Bureau of Statistics data. When comparable communal statistics are assembled for different years, data are said to form time-series. The Norwegian Commune Database contains both longitudinal statistics in the form of time-series and data that are incomparable in the sense that they pertain to single years only, depending on what information is available through the Bureau of Statistics.

The Database consists of two parts: 1) the data; 2) the programs that retrieve and manipulate the data. The data part comprises all kinds of official statistics such as data on demography, economy (both tax return and treasury data), industries, schools, elections etc.<sup>1</sup> In principle, the Database will include all published statistics at the commune level that social researchers might want to use. Since some data are more relevant than others, statistics are included according to a priority list. Currently, the Norwegian Social Science Data Services (NSD) has assembled in machine-readable form nearly 8,000 variables from the post-war period and 7,000 variables covering the first 100 years of communal rule from 1838 onwards.

The need for an aggregate-level, longitudinal database such as this one dates back to the late 1950s. The program of Norwegian electoral studies initiated at that time created a demand for various data linked to election statistics. This research program was developed by Henry Valen at the Institute for Social Research in Oslo in cooperation with Stein Rokkan at the Chr. Michelsen Institute in Bergen. Valen concentrated on the most recent data, while Rokkan started work on historical data<sup>2</sup>. In the late 1960s NSD was conceived as a data archiving institution, and the Commune Database was to become one of the main databanks of the Data Services<sup>3</sup>.

<sup>1</sup> Rokkan, Stein, and Henrichsen, Bjørn (eds.), *Kommunedatabanken. En håndbok for brukere*, NSD Handbook No. 2, Bergen 1977.

<sup>2</sup> Valen, Henry, and Rokkan, Stein, *The Norwegian Program of Electoral Research*, in: Pesonen, Pertti (ed.), *Scandinavian Political Studies*, Vol. 2, Helsinki, New York 1967.

<sup>3</sup> Rokkan, Stein, and Henrichsen, Bjørn, *Building Infrastructures for the Social Sciences: The Norwegian Social Science Data Services, Research in Norway*, Oslo 1976.



The activity of the Data Services during the first years of operation concentrated on programming a system for the management of communal data which at the time were divided into separate files of punched cards. One of the tasks of the Data Services was to organize the data so that they could be accessed in an efficient manner. Also, a computer program was designed to keep track of administrative boundary changes at the commune level.

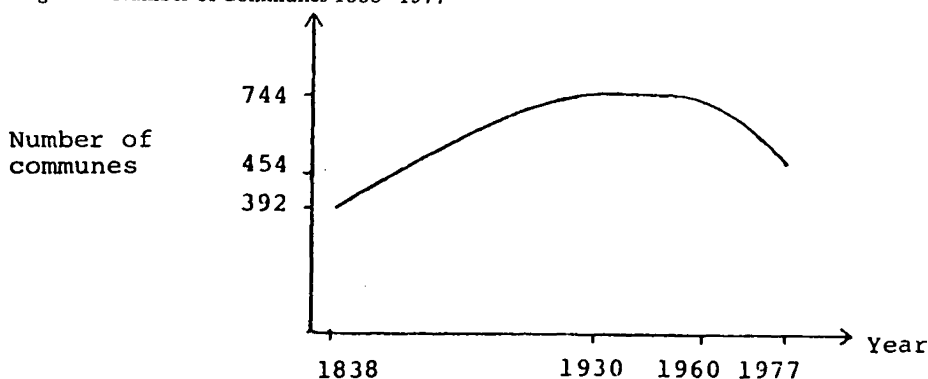
The Commune Database as a programming project has continued up to now. Part of the programs will soon be due for replacement by more efficient database management systems, as the old FORTRAN routines gradually get out of data. However, some features are unique, notably the standardization process which the Database system can carry out.

### Boundary Changes

When they were established in 1838, the law enabled communes to divide on their own initiative, which brought about an increasing number of new commune units. After World War II the role of the communes in the nation's economy meant that some kind of *arrondissement* or consolidation had to be instituted. Consequently, during the 1960s more than half the total number of communes were consolidated into larger administrative units<sup>4</sup>.

The number of communes, originally 392 in 1838, increased to 744 in 1930. From then on, only small changes took place until the consolidation phase in the 1960s and 1970s. As of writing there are only 454 communes. This development is shown graphically in Figure 1.

Figure 1: Number of Communes 1838–1977



<sup>4</sup> Lie, Suzanne, and Taylor, John G., A Sociological and Geographical Appraisal of Commune Boundary Changes in Norway, mimeo; Oslo 1977.

Figure 2: Nature and Effect of Border Changes

Nature of Change		Effect of Change	
		Creation of New Unit	Extension of Old Unit
		A	B
	Splits		
	Mergers	C	D

Border changes have the effect of creating new units or extending old ones. Also, the nature of change can be classified as splits or mergers, representing divisions and consolidations respectively. These two dimensions can be grouped in a four-fold table as shown in Figure 2. The table indicates that new units can be established and old ones extended either through split-ups or mergers.

Figure 2 forms a departure point for the coding of border changes. The changes are viewed as having two effects, one for the parting commune and one for the receiving commune. In this way border changes must be recorded bilaterally, and a separate code given to each of the units to indicate whether it is the receiving or parting commune. If more than two communes are involved in the same division or consolidation, a multiple of recorded changes will result in order to register each bilateral relation between parting and receiving communes separately.

The codes used in the Norwegian Commune Database are shown in Figure 3. Two codes are allotted to each of the cells of the table in Figure 2. These codes can be compared to the codes of the Swedish GEOKOD register except for the fact that until the early 1950s, no Swedish commune was extended by split-ups of other established units<sup>5</sup>.

Figure 3: Commune Database Border Change Codes

Categories cf. Fig. 2	Code	Explanation
A	1	x is separated from y
A	2	y parts with x
B	3	x incorporates part of y
B	4	y is split to extend x
C	5	x is transformed from y
C	6	y is transformed into x
D	7	x incorporates y
D	8	y is incorporated into x

<sup>5</sup> Öhngren, Bo (ed.), *Geokod. En kodlista för den administrativa indelningen i Sverige 1862–1951*, Uppsala 1977.

## Principles of Standardization

The communal border fluctuations complicate the problem of undertaking longitudinal studies at this level of aggregation. In order to apply frequently used statistical computer programs like SPSS to analyze time-series, data must be standardized to fit the ordinary „flat file“ data matrix which presupposes an equal number of variables for all units. The program routines of the Commune Database were designed to do this standardization as part of the retrieval functions invoked when the user requests a particular set of variables.

In principle, there are three ways to standardize a data matrix and take boundary changes into account without resorting to missing value codes indiscriminately to fill up the matrix. The three ways of eliminating troublesome boundary changes are as follows:

- (a) units can be dropped
- (b) units can be aggregated
- (c) properties can be estimated while keeping the actual number of units.

The simplest solution is to drop all units that have been modified. If the drop-out rate proves too high, units can be kept if the alterations involve certain amounts of the resources or significant communes. Still, this is an unsatisfactory solution due to the fact that important values or units might be excluded. If a device can be found to keep as much data as possible in the standardized file, there will be fewer risks of overlooking important variations in the data.

A more agreeable solution is to effect an aggregation of units that merge. The main drawback of this approach is that the process of aggregation will conceal information, making the measures less detailed than in the original file. Also, problems arise when units split and merge simultaneously, in which case a combination of aggregation and omission of units impair the geographical distribution of properties considerably.

The third possibility is to determine the extent to which resources are redistributed as a result of border changes. The fractions indicating resource transfers can be turned into *standardization coefficients* to be applied to data values when time-series analyses are carried out. The idea is that data values for the various units must be standardized as required to facilitate the combination of data from before and after each border change. The standardization process will be most suitable when official statistics on the direct effects of the changes can be found. More imprecise results can be obtained if data have to be estimated or computed. However, some kind of estimation procedure will be required whenever statistics pertaining to the changes are unavailable.

The last of these three solutions is the most complicated and requires a specially written computer program to be effective. The complications stem from the fact that in the extreme case, separate standardization coefficients will be available for

every property to be redistributed bilaterally. Obviously, the computational procedure can be simplified if more rudimentary coefficients are relied upon. Sometimes, reliance on rudimentary coefficients is required due to the huge amount of work involved in setting up a complete series of standardization coefficients, or coefficients could be partly unavailable due to incomplete statistics.

In Norway, statistics on the direct effects of the border changes are sadly missing. In effect, only two kinds of information are available, namely the size of the area divided or consolidated, and the number of individuals concerned. Size of the area is of little interest in the social sciences, which leaves us with the population figure as our only point of departure, if other measures cannot be estimated.

In the face of the unsatisfactory solution of dropping or aggregating units, locally produced estimates can be relied on to provide coefficients for the standardization process when statistics on the direct effects are unavailable. Theoretically, this strategy constitutes a feasible solution. However, in Norway the idea was impossible to realize in a reasonable way because of the size of the problem. Since 1838 there have been more than 1,000 border changes. Consequently, the number of variables for which standardization coefficients must be found, would be very high. In the end it appeared that we had to rely on a computational procedure based on the available population figures to estimate simplified standardization coefficients.

No doubt a certain amount of error is introduced when a single measure such as the amount of population transferred is used as a redistribution criterion. As always, it is questionable whether resources forming time-series will be correctly redistributed by application of a time-specific coefficient. Also, the population distribution of a particular area has a specific geographic pattern, as people tend to group in agglomerations located close to rivers or fjords. What matters is that the settlement pattern will correlate to a greater or lesser extent with the location of other resources in the same area. Thus the population density can to some extent be used to deduce the geographical distribution of other resources.

Generally, densely populated areas are thought of as being better off in many respects compared to the scattered population of the periphery. To a certain extent measures describing economic well-being and material resources are related to the size and density of the population, whereas the dispersion of natural resources are not. In this way the computational redistribution will sometimes be fairly correct, sometimes wildly misleading, depending on the variables in question.

The bilateral transfer of resources can also be affected by differing resource profiles that characterize the communes. We are now thinking of extremes such as highland vs. lowland, rural vs. urban, inland vs. coastal, etc. Conditions such as these can affect the correctness of the estimated redistribution for instance if one of the communes is predominantly rural showing abundant natural resources, and the other one is predominantly urban, showing a striking amount of welfare facilities.

Also, the population of a commune can be unevenly scattered throughout the area so as to further upset the relationship between people and resources. If a certain percent of the population is ceded when part of one commune is transferred to a neighbouring commune, the transfer can involve the most atypical part of the

population as far as resources are concerned. More often than not transfers affect such atypical groups, as changes take place to improve the homogeneity and internal resource distribution of the communes.

Traditionally, the distribution of population and resources are unequal in Norwegian communes, so many of the adverse comments that we have made, will apply<sup>6</sup>. In spite of these apparently unfavourable conditions, the computational approach was selected for implementation. It remains to be seen how this approach affects the data, as no analysis has been carried out yet to evaluate the effects on time-series analysis. However, for the time being the simple computational approach seems best, as the dropping of units is deemed more erroneous, aggregation is partly prevented by concurrent split-ups and consolidations, and the summoning of local expertise to produce individual coefficients is too expensive and unreliable to be of much use.

## The Standardization Process

Even if careful estimates are made for every border change, some resources will be difficult to redistribute in a sensible way. Naturally, a crude computational procedure will to some extent give wrong results unless precautions are taken. The problems of the described heterogeneous distribution of people and resources cannot be fully resolved, but some of the undesirable effects can be removed.

Clearly, buildings and institutions cannot split. Neither can measures such as per capita income and public expenditure. Generally, when there are few members in a class of objects, or if the resource has no obvious relation to the population distribution, values cannot split between parting and receiving communes by being computed. Conversely, resources which appear to have a more continuous or homogeneous character throughout the population can be redistributed according to the amount of population transferred. Demographic variables lend themselves most favourably to such treatment and can be redistributed by means of the population figure almost unconditionally.

Unfortunately, it appeared that most of the variables in the Commune Database belong to the unsplitable category and had to be assigned missing values for both parting and receiving communes. To decrease the amount of missing values, *computability codes* were introduced. A computability code is associated with each variable to tell the computer program what to do when a pair of communes is subject to modifications.

<sup>6</sup> Lie and Taylor, op. cit.

The computability codes used in the Commune Database to indicate the computational measures to be taken, are as follows:

- (a) proportional computation
- (b) computation for consolidations only
- (c) transfer if values are identical numerically
- (d) no computation or transfer

Proportional computation means that the percent of total population transferred will be the coefficients used by the program to redistribute the variable values for the communes in question. Computation in the case of consolidations means that a summation of values will take place only for border changes belonging to categories C and D in Figure 2. These are cases of 100 percent transfer of resources because the parting unit ceases to exist. Computation only if the values of the parting and receiving communes are identical in numerical terms means that both parting and receiving communes retain their values unchanged. No computation means that missing data will be filled in for both parting and receiving communes, as no computation, transfer or summation of values can take place.

Each variable in the Database is assigned only one computability code upon judgement of which code is most suitable throughout the country. No local conditions that represent exceptions will be considered by the computer program during processing. Neither are the cumulative effects of successive changes taken into account. In this way there will be no check on the admittedly infrequent cases when a split neutralizes a previous merger or vice versa, as the system has no way of knowing if part of a commune has been affected twice.

The change codes explained in Figure 3 are only used for setting up the standardization coefficients. During actual processing these coefficients plus the computability codes constitute the constants relied on by the program to produce the standardized matrix. Figure 4 will give you an idea of the information that the program reads to enable the standardization process. First, the program will read the commune identification numbers and the information on years of establishment and discontinuation as indicated in Figure 4a. This information helps the program determine the actual number of communes for any single year. Next, information on border changes are accessed as specified in Figure 4b.

The data part of the Database is stored in a so-called transposed matrix (i. e., variablewise instead of the ordinary casewise structure) using the exact number of communes that existed in the year of data collection. This information is available to the program as part of the variable label. By means of the establishment file the program will identify which data elements belong to which communes, as the ordering of the data elements corresponds to the sequential order of the commune identification numbers.

Next, the program will read which year the user has specified as the year of standardization, as well as the list of variables that the user wants to extract from the Database. By means of the border change file and the computability codes the program will now standardize the data by applying the coefficients year by year from the year of data collection till the year of standardization. When all variables

**Figure 4a: File of Establishment and Discontinuation**

Commune identification number	Year of discontinuation	Year of establishment (default is 1838)
0101		
0102		
0103		
0104		
0111		
0112	1964	
.		
.		
.		
0402	1964	1965
.		
.		
.		
1201		1972
.		
.		
2030		

**Figure 4b: File of Border Changes**

Parting commune	Year	No. of receiving <sub>1</sub>	Coefficient <sub>1</sub>	Receiving commune <sub>1</sub>	Coefficient <sub>2</sub>	Receiving commune <sub>2</sub>
1264	1945	1	28.335	1265		
0218	1947	1	100.000	0301		
1041	1948	1	1.542	1003		
0814	1949	1	2.661	0802		
.						
.						
.						
0537	1962	2	8.383	0538	91.616	0536
.						
.						
.						
1941	1974	1	31.884	1942		
1855	1974	1	100.000	1805		

have been treated in this manner, the user will have the data transposed back into the rectangular flat file structure ready for analysis by SPSS or any other statistical program package<sup>7</sup>.

## Further Capabilities of the Database

The Commune Database was originally thought of as a bank of post-war data from Bureau of Statistics publications and magnetic tapes. The supporting computer routines were meant to offer a solution of the standardization problem for the post-war period only, so they were not written with a general purpose in mind.

The most serious restriction of the original routines was the ability to standardize only forward in time. This solution was chosen to minimize the various errors that can be inflicted by the computational process. Most significantly, the post-war period brought about a majority of consolidations, as indicated in Figure 1. The consolidations resulted in an increasing number of larger and more resource rich communes as we approach the 1970s. Standardizing backwards in time would split these resource rich units into several relatively impoverished communes on false assumptions, as the heterogeneous character of the communes before consolidation cannot be deduced. The extreme case happens when a city has been merged with neighbouring communes that are predominantly rural.

Things become more complicated as the Database is extended back to 1838. Now the user must be able to choose whether standardization should be effected forwards or backwards in time. If all data belong to the period after 1930, standardization forwards will give the least errors, as there is a majority of consolidations in this period. Data belonging entirely to the period before 1960 should be standardized backwards because of the divisions<sup>8</sup>. This is a useful rule of thumb which can be trusted in most cases. If the user wants to combine pre-1930 with post-1960 data, no advice can be given. In this case a close scrutiny of modifications and variables is required to see how standardization affects the data.

Because of the extended time-span of the present Database, some additional requirements must be satisfied to have an acceptable retrieval system for time-series data of nearly 150 years. One of the most important improvements is a report facility that users can involve to get a modification report on paper. This modification report will consist of a listing in tabular form of the variable values before and after

<sup>7</sup> Brosveet, Jarle, Teknisk dokumentasjon for Kommunedatabanken 1945–1977, NSD Report No. 15, Bergen 1977.

<sup>8</sup> Aarebrot, Frank H., and Kristiansen, Bjarne, Norwegian Ecological Data 1868–1903, NSD Report No. 9, Bergen 1976.



standardization. By consulting the list, the user can inspect changes to see if corrections are needed.

Users also want to control other parts of the standardization process, such as the coefficients and the computability codes to be used, as the decisions made at the time of implementation will not suit all needs. Though the user is discouraged from meddling with the computational process, errors can be corrected more efficiently by specifying an alternative set of coefficients and codes instead of updating values in the standardized file.

A seemingly dispensable improvement is the selection feature that enables the user to have a subset of standardized communes on file. However, it has proved important that the standardized file produced by the Database is as ready for analysis as possible. Even if selection routines are found in most statistical program packages, users find it tedious to start an analysis run by filtering out units they never wanted from the Database.

Another requirement is the need to add data temporarily. This requirement stems from the wish of users to include data that are not part of the Database. In this way the user can have his own data standardized and included on the same file as any Commune Database variables that he retrieves. User-supplied data may be derivatives or raw data from sources other than the Bureau of Statistics that NSD does not include in the regular updating of the Database. Naturally, if user-added data are of sufficient general interest, they will be included permanently, if no clauses are attached.

## Future Development

The Commune Database is but one of the conceivable systems for the retrieval of locality data. A host of other units at various levels also need attention. Separately, NSD has made available a Census Tract Database for 1960 and 1970 data aggregated from person level records. To some extent the Census Tract Database takes care of the linkage to the Commune Database, though its primary objective is to standardize data at the census tract level<sup>9</sup>.

It is hardly feasible to build separate systems for each locality level, so some kind of general solution is called for. No doubt, experience from the Commune Database project and the Census Tract Database project will offer some guidance on how to design a generalized linkage system for aggregate data. In particular, a

<sup>9</sup> Alvheim, Atle, NSDs Kretsdatabank. Datainnhold og brukerveiledning, NSD Report No. 18, Bergen 1977.

general linkage system will be needed because of the difficulties of drawing definite conclusions based on data for only one level of aggregation.

Another major development trend is the setting up of a Cartographic Database that will enable data from the Commune Database to be drawn as maps<sup>10</sup>. Since the geographical distribution of variable values often appears to form striking patterns, maps will be the best way of discovering regional differences. Statistical graphics can enhance the display of geographic patterns even more, so a graphics system coupled to the map-drawing routines is being tested<sup>11</sup>.

The utility of the Database can be improved even further if data were available at the individual level. The 19th century censuses will be available soon as they can be released without violating personal integrity as protected by the Person Register Law. Modern data are harder to get hold of, as most personal and locality identifications are likely to be removed from the data when they are made available to the Data Services. However, academically conducted surveys and background information on public figures do not suffer from this restriction, although such data are mostly too specific to be of general interest.

The Norwegian Social Science Data Services will also store specific interest data as required, but as far as the Commune Database is concerned, most potent research opportunities will result when census data are combined with communal time-series. NSD will take steps to improve the linkage of such data as one of our major projects for the 1980s.

<sup>10</sup> Sande, Terje, Cartographic Database for Time-Variable Data, EPD 25 (December 1977), a contributed paper to the European Meeting on Regional Databases for Computer Cartography, Bergen 1977.

<sup>11</sup> Alvheim, Atle, Figur: A Program for Point Graphics, EPD 25 (December 1977), a contributed paper to the European Meeting on Regional Databases for Computer Cartography, Bergen 1977.

## HIWED – A Comparative Historical Research Project on Western Europe

The main purpose of this contribution is to report on a comparative-historical research project, i. e. to define our basic objectives and to describe our approach in pursuing them. It does not attempt to discuss more specific theoretical questions or methodological problems investigated in the project or to present empirical results. The intention is to give the reader both a general impression of the project in order to stimulate general curiosity, and enough detailed information to create more specific interest.

„HIWED“ is the acronym for „*Historical Indicators of the Western European Democracies*“. This is the name of a macrosociological research project which was started in Fall 1973 by Wolfgang Zapf and myself at the University of Mannheim. Since Spring 1977 it has been directed by myself at the Forschungsinstitut für Soziologie of the University of Cologne. The staff consists of three sociologists: Jens Alber, Jürgen Kohl, and Winfried Pfenning, two economists: Franz Kraus and Kurt Seeböhm, and one political scientist: Richard Eichenberg. The project is generously funded by the Stiftung Volkswagenwerk through the Fall of 1979.

The HIWED-project has two main objectives<sup>1</sup>. The first involves compilation of a historical data handbook for all Western European countries (with the exception of Greece, Portugal and Spain) for the time period 1815–1975. This task was our main concern in the first phase of the project conducted between 1973 and 1976. In the second three-year phase of the project – with a greatly expanded staff – the emphasis shifted toward the second goal, namely the attempt to analyze various aspects of the historical development of welfare states in Western Europe in comparative perspective and within a more general theoretical frame.

This objectives are closely related and support each other, but they also restrict one another. A larger part of the data compiled for the handbook are being used for description, analysis and „explanation“ of welfare state development, and thereby the reliability and comparability of the respective data have probably been improved. On the other hand, however, the phenomena and questions which can be analyzed

<sup>1</sup> For a more detailed discussion of these two objectives cf.: Flora, P., *The HIWED Project – The Handbook, Theoretical Orientations and Statistical Sources*. HIWED-Report No. 1, Mannheim 1975, and Flora, P., *Das HIWED-Projekt – Ein Zwischenbericht*, pp. 318–340, in: Zapf, Wolfgang (ed.), *Soziale Indikatoren. Konzepte und Forschungsansätze*, Vol. 3, Frankfurt/M. 1975.

in a comparative perspective are somewhat limited by the kinds of data: only quantitative aggregate data and institutional data are suitable for a more systematic compilation in a comparative historical handbook.

The project is greatly inspired by Stein Rokkan whose work represents the most important revival of Max Weber's comparative historical macro-sociology — a revival in a changed professional context which both requires and enables more systematic empirical research<sup>2</sup>. Rokkan persistently encourages a comparative approach which had been axiomatic for classical sociology but which for a long period had lost its centrality. For Durkheim it was obvious that „la méthode comparative est la seule qui convienne à la sociologie“. For macro-sociological questions, most obviously, this implies comparison over time and space; and the area of European history is probably the best opportunity for such comparisons.

## A Historical Data Handbook for Western Europe

Our effort to produce a historical data handbook for Western Europe should be seen in the context of other attempts to establish a more systematic data basis for the analysis of problems and developments of total societies. Between the middle of the nineteenth century and the First World War a great variety of comparative historical data collections were compiled creating a strong and rich tradition. The roots of this tradition reached back to the English Political Arithmetic with its exclusive emphasis on quantitative data, on the one hand, and to the German Universitätsstatistik with a greater interest in institutions, on the other. The interwar period produced a break in this tradition — coincident with, though not directly related to the break in the classical macro-sociological tradition — which was not renewed until around 1960<sup>3</sup>.

The early 1960's saw a first wave of cross-national data collections. The most important of these are Ginsburg's *Atlas of Economic Development, A Cross-Polity Survey* by Banks and Textor, and the first *World Handbook of Political and Social Indicators* by Russett et al., later followed by a second enlarged edition by Taylor and Hudson<sup>4</sup>.

<sup>2</sup> An attempt to „summarize“ the historical sociological work of Stein Rokkan can be found in the HIWED-Report No. 1.

<sup>3</sup> A more detailed description of these developments is given in: Flora, P., *A New Stage of Political Arithmetic*, in: *Conflict Resolution* 18, No. 1 (1974), pp. 143–165.

<sup>4</sup> Ginsburg, N., *Atlas of Economic Development*, Chicago 1961; Banks, A. S., and Textor, R. B., *A Cross-Polity Survey*, Cambridge/Mass. 1963; Russett, B. M., et al., *World Handbook of Political and Social Indicators*, New Haven and London 1964; Taylor, C. L., and Hudson, M. C., *World Handbook of Political and Social Indicators*, 2nd ed., New Haven and London 1972.

In the second half of the 1960's increased effort was made to compile comparative historical data collections. Among the most important projects are<sup>5</sup>: the „Correlates of War Project“ of Singer and Small at the Mental Health Research Institute, the University of Michigan; the „Minnesota Political Data Archive“ of Flanigan and Fogelman at the University of Minnesota, the „Suny-Binghamton Cross-National Time-Series Data Archive“ of Banks at the State University of New York, Binghamton; the project on „Comparative Causes of Societal Stability and Instability“ by Hage at the University of Wisconsin; the „QUAM-Project“ (Quantitative Analysis of Modernization) of Zapf and Flora at the University of Frankfurt; and the „Nordic Countries Project“ by Kuhnle at the University of Bergen.

The major comparative historical data collections published in this second wave are *The Wages of War, 1816–1965* by Singer and Small, *Cross-Polity Time-Series Data* by Banks, and most recently *European Historical Statistics, 1750–1970* by Mitchell<sup>6</sup>.

The QUAM-project<sup>7</sup> conducted by Wolfgang Zapf and myself was to some extent a pre-stage of the HIWED-project, although, of course, only in retrospect. In this earlier and much smaller project we produced various data collections which were presented in 1971 at the „ISSC/ECPR Workshop on Indicators of National Development“, organized by Stein Rokkan at the University of Lausanne. This workshop brought together for the first time major projects with similar objectives and gave a strong impetus to future work.

It encouraged us to start in 1973 the HIWED-project with the intention of producing a handbook on Western Europe with more reliable and comparable (which is always a matter of degree) „indicators“ of major social, economic and political development, including growth processes as well as structural changes. The handbook will mainly consist of time-series of national aggregate data, supplemented by qualitative indicators of institutional changes. The time period covered will vary, embracing at least the years between 1920 and 1975, and a maximum time span from 1815 to 1975. The handbook will be strictly comparative. That is, it will contain only data which we were able to compile for all (or almost all) Western European countries (with the exception of Greece, Portugal and Spain). It will be finished in Spring 1979 in the form of two volumes of approximate 500 pages each. The table of contents on the two following pages will give a first impression of the topics included in the handbook.

<sup>5</sup> Short description of these projects can be found in: Flora, P., *Quantitative Historical Sociology*, in: *Current Sociology*, 2 (1975).

<sup>6</sup> Singer, J. D., and Small, M., *The Wages of War 1816–1965*, New York 1972; Banks, A. S., *Cross-Polity Time Series Data*, Cambridge/Mass. 1971; Mitchell, B. R., *European Historical Statistics 1750–1970*, London 1975.

<sup>7</sup> Cf.: Zapf, W., *Materialien zur Theorie des sozialen Wandels*, Konstanz 1967; Zapf, W., and Flora, P., *Some Problems of Time-Series Analysis in Research on Modernization*, in: *Social Science Information*, 3 (1971), pp. 53–102; Flora, P., *Modernisierungsforschung – Zur empirischen Analyse der gesellschaftlichen Entwicklung*, Opladen 1974; Flora, P., *Indikatoren der Modernisierung – Ein historisches Datenhandbuch*, Opladen 1975.

# State, Economy, and Society in Western Europe

A data handbook 1815–1975

## Volume I

### The Growth of Mass Democracies and Welfare States

#### *A Boundaries*

1. Territorial change
2. Cultural heterogeneity

#### *B Mass Democracies*

3. Elections and parties
4. Parliaments and coalitions

#### *C Resources of the State*

5. Public revenues
6. Public expenditures

#### *D Personnel of the State*

7. Public bureaucracies
8. Military

#### *E Welfare and Education*

9. Social security
10. Literacy and education

## Volume II

### The Growth of Industrial Societies

#### *A Population and Families*

1. Population growth
2. Families

#### *B Urbanization and Housing*

3. Cities and urbanization
4. Housing conditions

#### *C Economic Growth*

5. National Product: aggregates
6. National Product: origin and use

#### *D Division of labour and inequality*

##### 7. Labour force

##### 8. Income distribution

#### *E Trade unions and Strikes*

##### 9. Trade unions

##### 10. Strikes

Several steps in the production of the handbook can be distinguished:

(1) Our first task involved reconstructing the history of official statistics and compiling a bibliography of official statistics for all Western European countries since the beginning of the nineteenth century, a task which — to our surprise — the history profession has neglected until recently. A comparative history and bibliography of national official, international and private statistics was finished in early 1975 and was published in *Current Sociology* in my trend report on „Quantitative Historical Sociology“<sup>8</sup>. Since then we have also finished a bibliography of all census publications in Europe since 1800. Additional bibliographies of more specific statistical publications (e. g. of public revenues) are in preparation and we may decide to integrate the various bibliographies into one book.

(2) We also started a data archive, which consists (or will consist) at the end of the project in 1979) of: a) microfiche of all statistical yearbooks from the beginning to 1965, b) photo-copies of all published census results since 1801 (but for the national level only and somewhat selected), c) copies of the main publications of official statistics for a variety of fields (public finance and bureaucracies, social security, income distribution, education, strikes, demography), d) copies of the main publications of international statistics until World War II, e) copies of the most important „private“ statistical collections (comparative and old national collections), f) copies of a great variety of important articles from statistical journals.

To my knowledge there will be no single library in Europe offering the same easy access and potential for systematic comparative study of societal development in Western Europe using official statistics. Therefore, it will be necessary to think about „institutionalization“ of the archive after the end of the HIWED-project in order to make it a useful instrument for future comparative research.

(3) A third step involves production of data reports on specific topics which will become chapters of the handbook. These reports contain much more data than will be included in the handbook, and they give complete documentation (i. e. in principle each single number is documented). A distinction is made between a first part with „country tables“ and a second part with „comparative tables“ in order to make the data strictly reproducible on the one hand, and relatively comparable, on the other. In addition, each report contains an introduction discussing problems of reliability and comparability as well as a bibliography. Data reports on the follow-

<sup>8</sup> See note 5. This issue of *Current Sociology* also contains our first data collection in form of 21 tables.

ing topics have been finished or will be finished soon: population, urbanization, labor force, national product, income distribution, public bureaucracies, revenues and expenditures, social security, housing, trade unions, strikes, and elections.

We hope that these data reports can be made available in the form of micro-fiche. They would be used for documentation of the computerized data sets which we began processing in early 1978 with the very generous help of the Historical Archive of the ICPSR in Ann Arbor and the Zentralarchiv at the University of Cologne. Thus, we hope that with the end of the HIWED-project all data contained in the reports will be available on tape in combination with extensive documentation on micro-fiches.

## On the Development of the Western European Welfare States

The search for the statistical sources, building the archive, and collecting data have been our main tasks during the first phase of the project simply because they are extremely time-consuming. Nevertheless, we have also consistently pursued a second and more substantial goal: analysis of the growth of welfare states in Western Europe covering a variety of different aspects on that growth. Although the extension of the archive still requires great expenditure of time, these studies have now become our major task.

It was not primarily specific concern with social policies which shaped this goal, but rather an interest in Stein Rokkan's theory of European political development. This theory essentially consists of two parts: a conceptual-analytical model on the one hand, and empirical typologies of macro-configurations, on the other. The conceptual-analytical model embraces:

(a) a conception of basic structures which combines Parsons' functional distinction between four societal subsystems with a 'hierarchical' distinction between center (res) and periphery(ies), and

(b) a corresponding conception of four basic developmental processes or problems the respective predominance of which may lead to specific stages or even crises:

(1) *State-Formation*: the formation of fiscal and military states; the phase (problem) of political, economic and cultural unification at the elite level, of creation of organizations for the mobilization of resources (tax-bureaucracies), for consolidation of the territory (armies), and for maintenance of internal order (police and army).

(2) *Nation-Building*: the building or growth of national states; the phase (problem) of establishment of direct contacts between the elite and larger sectors of the peripheral population through conscript armies, schools and mass media, of religious and linguistic standardization, and of development of feelings of national identity.

(3) *Participation*: the development of mass democracies; the phase (problem) of



establishment of political citizenship through the equalization of political rights: involving growing participation of the dominated population, the institutionalization of civil and political rights (franchise, parliaments), and creation of political parties.

(4) *Redistribution*: the development of welfare states; establishment of social citizenship through the redistribution of resources, goods and benefits: through the creation of public welfare systems (social security, health, education, housing) and of public policies for the equalization of economic conditions through progressive taxation and transfer payments.

Rokkan has devised several empirical typologies of macro-configurations specifying the general model in the context of Western European history. The main typologies look toward explanation of variations in territorial consolidation, the relationships between external consolidation and internal restructuring, the introduction and extension of the franchise, and the development of cleavage structures and party systems in Europe. His basic approach involves explaining variations of macro-configurations at a specific point of time by earlier macro-configurations *plus* new developments or events (e. g. the Reformation or industrialization). Progress in theory-building thus implies a persistent effort to improve the typologies of macro-configurations, to invent new ones and to combine them into a consistent whole.

Rokkan has not yet incorporated the fourth „phase“, the growth of welfare states, into his typological model. This has stimulated our attempt to analyze the development and variations of welfare institutions in Western Europe and to relate them to earlier or concomitant developments and variations in terms of Rokkan's typologies. The general question is how different developments in the first three „phases“ have created macro-settings promoting or retarding the development of welfare states. A first result of these attempts was presented at the IPSA World Congress at Edinburgh in 1976 in our paper: „On the Development of the Western European Welfare States“<sup>9</sup>.

Working with Rokkan's stimulating approach, we soon encountered two related theoretical problems:

(a) his approach emphasizes analysis and explanation of variations rather than uniformities, and the development of welfare states in Western Europe shows uniform features to an important degree;

(b) it is an approach which focuses upon the political collectivity or more generally – in Weber's sense – on the „Verbandsstruktur“ of societies, relatively neglecting socio-economic and sociocultural developments or introducing them as exogenous variables.

In his emphasis on historical variety (including a certain distrust of evolutionary theory) and his accentuation of „*politics* between economy and culture“, Rokkan's approach is essentially Weberian. In order to deal with the two limitations mentioned above, it will be necessary to incorporate theoretical elements of other macro-

<sup>9</sup> This paper has now been published in German: Flora, P., et al., Zur Entwicklung der west-europäischen Wohlfahrtsstaaten, in: Politische Vierteljahresschrift, 4 (1977), pp. 707–772.

sociological traditions — a task which has not yet been attempted in the analysis of the welfare state. In our view, the two most relevant traditions here are those of Marx and Durkheim, emphasizing standardizing socio-economic and socio-cultural developments respectively.

The Marxian perspective suggests that the effort to understand the creation of the (relatively uniform) institutions of the welfare state is an attempt to deal with the (relatively uniform) socio-political problems (economic crises and class conflict) created by the dynamics of a relatively autonomous economic system. The Durkheimian perspective, on the other hand, suggests that the effort to understand the creation of the (relatively uniform) institutions of the welfare state is an attempt to solve the (relatively uniform) socio-cultural problems („moral“ integration, social solidarity) created by an ongoing process of functional differentiation<sup>10</sup>.

Whereas these two theoretical traditions thus help to understand primarily the general emergence of the welfare state as a response to common developmental problems, Rokkan's theory will serve to explain the divergent institutional solutions found for these problems. The attempt to integrate the various theoretical perspectives into a more general theory of the development of the welfare states is mainly motivated by the requirement of developing a common framework for the empirical studies which represent the true core of the project. These studies deal with specific aspects of the growth of welfare states in a comparative and historical perspective using quantitative aggregate data as well as qualitative institutional data. The aspects analyzed by the individual members of the project are:

Social security

by Jens Alber<sup>11</sup>

Public bureaucracies

by Richard Eichenberg<sup>12</sup>

<sup>10</sup> A first attempt to incorporate some of Durkheim's ideas into a macro-sociological study of the welfare state was made by Peter Flora in a paper „Durkheim and the Welfare State“ presented in July 1977 at a workshop on „The Western Welfare States: Historical Developments and Current Problems“ organized by the HIWED-group and the Social Policy Group of the Committee on European Studies.

<sup>11</sup> Alber has produced a data report on the coverage of the four major social insurance systems in Western Europe since their beginnings, containing quantitative as well as institutional data; a first analysis of the institutional data was carried out in our Edinburgh paper — cf. note 9; furthermore, he has studied the development of unemployment insurance systems on the basis of the standardized detailed characteristics of all major unemployment insurance laws; an analysis of these laws was presented at the above mentioned Luzern-workshop under the title „Governments, Unemployment and Unemployment Protection: On the Development of Unemployment Insurance in Western Europe“. A revised and extended version of this study has been presented at the 9th World Congress of Sociology at Uppsala.

<sup>12</sup> Eichenberg, who joined us as a computer specialist coming from Ann Arbor, Michigan, started a first analysis at the 1979 APSA meetings in New York under the title, „The Growth of Public Bureaucracy in Five European Countries“. The data report should be finished in 1979.

Public expenditures  
Income distribution  
Trade unions and strikes  
Public revenues

by Jürgen Kohl<sup>13</sup>  
by Franz Kraus<sup>14</sup>  
by Winfried Pfenning<sup>15</sup>  
by Kurt Seeböhm<sup>16</sup>

Of these six theses<sup>17</sup>, three will be finished by the end of the project (Alber, Kohl, Kraus), the other three will require additional effort. Besides the data handbook, they will represent the substantial product of the HIWED-project marking its relative success or failure.

<sup>13</sup> Kohl has finished a data report on enfranchisement and electoral participation, but his major subject is a study of the development of public expenditures; a first analysis can be found in our Edinburgh paper — cf. note 9; at the Luzern workshop he presented an analysis of the growth and structural change of public expenditures in Western Europe after World War II which soon will be extended to earlier periods.

<sup>14</sup> Kraus has finished a data report on national product and is completing his collection of income distribution statistics until the end of 1978; he has done a critical study of the historical tax statistics as well as of the major studies using such statistics; this study was presented at the Luzern workshop under the title „Income Distribution in Western Europe: How Inequalities Can Be Assessed“.

<sup>15</sup> Pfenning has finished in 1977 a Diplomarbeit on „Gewerkschaften und Streiks in Westeuropa 1880–1975“, including data reports on strikes and trade unions; he will continue this work.

<sup>16</sup> Seeböhm has finished in 1976 a Diplomarbeit on „Die Entwicklung der Steuerstruktur im Prozeß der Modernisierung“, including a data report on taxes; he continues his work, with special emphasis on income taxes; together with Flora he presented a paper on the „European Tax Systems in the Welfare State Era“ at the Luzern-workshop.

<sup>17</sup> In addition a Diplomarbeit should be mentioned by Elisabeth Gransche on the „Wohnungsverhältnisse und Wohnungspolitik in Westeuropa — ein historischer Vergleich“ which was finished in 1976.